

Privacy Aware Learning

John C. Duchi¹ Michael I. Jordan^{1,2} Martin J. Wainwright^{1,2}

{jduchi,jordan,wainwrig}@eecs.berkeley.edu

¹Department of Electrical Engineering and Computer Science

²Department of Statistics
University of California, Berkeley

October 2012

Abstract

We study statistical risk minimization problems under a privacy model in which the data is kept confidential even from the learner. In this local privacy framework, we establish sharp upper and lower bounds on the convergence rates of statistical estimation procedures. As a consequence, we exhibit a precise tradeoff between the amount of privacy the data preserves and the utility, as measured by convergence rate, of any statistical estimator or learning procedure.

1 Introduction

There are natural tensions between learning and privacy that arise whenever a learner must aggregate data across multiple individuals. The learner wishes to make optimal use of each data point, but the providers of the data may wish to limit detailed exposure, either to the learner or to other individuals. It is of great interest to characterize such tensions in the form of quantitative tradeoffs that can be both part of the public discourse surrounding the design of systems that learn from data and can be employed as controllable degrees of freedom whenever such a system is deployed.

In this paper we approach this problem from the point of view of statistical decision theory. The decision-theoretic perspective offers a number of advantages. First, the use of loss functions and risk functions provides a compelling formal foundation for defining “learning,” one that dates back to Wald [42] in the 1930’s, and which has seen continued development in the context of research on machine learning over the past two decades. Second, by formulating the goals of a learning system in terms of loss functions, we make it possible for individuals to assess whether the goals of a learning system align with their own personal utility, and thereby determine the extent to which they are willing to sacrifice some privacy. Third, an appeal to decision theory permits abstraction over the details of specific learning procedures, providing (under certain conditions) minimax lower bounds that apply to any specific procedure. Fourth, the use of loss functions, in particular convex loss functions, in the design of a learning system allows the powerful tools of optimization theory to be brought to bear. Not only are optimization-based learning systems often successful in practice, but they are also often amenable to theoretical analysis. Finally, the decision-theoretic framework is a probabilistic framework, bringing probabilities to bear in the transformation from losses to risks, and this provides a natural hook for the use of randomization to provide control over privacy.

In more formal detail, our framework is as follows. Given a compact convex set $\Theta \subset \mathbb{R}^d$, we wish to find a parameter value $\theta \in \Theta$ achieving good average performance under a loss function $\ell : \mathcal{X} \times \mathbb{R}^d \rightarrow \mathbb{R}_+$. Here the value $\ell(X, \theta)$ measures the performance of the parameter vector $\theta \in \Theta$

on the sample $X \in \mathcal{X}$, and $\ell(x, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is convex for $x \in \mathcal{X}$. We measure the expected performance of $\theta \in \Theta$ via the risk function

$$\theta \mapsto R(\theta) := \mathbb{E}_P[\ell(X, \theta)], \quad (1)$$

where the expectation is taken over some unknown distribution P over the space \mathcal{X} .

In the standard formulation of statistical risk minimization, a method \mathcal{M} is given n samples X_1, \dots, X_n , each drawn independently from P , and its goal is to output an estimate $\hat{\theta}_n$ that approximately minimizes the risk function R . In this paper, instead of providing the method \mathcal{M} with access to the samples X_1, \dots, X_n , however, we study the effect of giving only some disguised view Z_i of each datum X_i . With $\hat{\theta}_n$ now denoting an estimator based on the perturbed samples Z_i , we explicitly quantify the rate of convergence of $R(\hat{\theta}_n)$ to $\inf_{\theta \in \Theta} R(\theta)$ as a function of the number of samples n and the amount of privacy provided by Z_i .

1.1 Prior work

There is a long history of research at the intersection of privacy and statistics, going back at least to the 1960s, when Warner [43] suggested privacy-preserving methods for survey sampling, and to later work related to census taking and presentation of tabular data (e.g., [19]). More recently, there has been a large amount of computationally-oriented work on privacy [17, 15, 47, 44, 22, 11, 5, 7, 38]. We overview some of the key ideas in this section, but cannot hope to do justice to the large body of related work, referring the reader to the comprehensive survey by Dwork [15] and the statistical treatment by Wasserman and Zhou [44] for background and references.

Most work on privacy attempts to limit disclosure risk: the probability that some adversary can link a released record to a particular member of the population or identify that someone belongs to a dataset that generates a statistic [13, 14, 37, 26]. In the statistical literature, work on disclosure limitation and so-called linkage risk, for example as in the framework of Duncan and Lambert [13, 14], has yielded several techniques for maintaining privacy, such as aggregation, swapping features or responses among different datums, or perturbation of data. Other authors have proposed measures for measuring utility of released data (e.g., [26, 9]). The currently standard measure of privacy is differential privacy, due to Dwork et al. [17], which roughly states that $\hat{\theta}_n$ must not depend too much on the n samples, and it should be difficult to ascertain whether a vector x belongs to the set $\{X_1, \dots, X_n\}$ given $\hat{\theta}_n$. Formally, paraphrasing the definition of Wasserman and Zhou [44], the method \mathcal{M} has α -differential privacy if

$$\sup_{S \in \sigma(\Theta)} \sup_{x_1, \dots, x_n} \sup_{x'_1, \dots, x'_n} \frac{\mu(S \mid X_1 = x_1, \dots, X_n = x_n)}{\mu(S \mid X_1 = x'_1, \dots, X_n = x'_n)} \leq \exp(\alpha). \quad (2)$$

where the sets x_1, \dots, x_n and x'_1, \dots, x'_n differ in at most one element, $\mu(\cdot \mid X_1, \dots, X_n)$ is (a version of) the conditional probability of the estimator $\hat{\theta}$ constructed by the method \mathcal{M} using the n samples, and $\sigma(\Theta)$ is a suitable σ -algebra on Θ .

Differentially private algorithms enjoy many desirable properties [17, 15] and essentially guarantee that even if an adversary knows all the entries in a dataset but the n th, it is difficult to discern whether a vector x is equal to X_n given the output of the method \mathcal{M} . Several researchers have studied differentially private algorithms for empirical risk minimization, providing guarantees on the excess risk of differentially private estimators $\hat{\theta}$. Chaudhuri et al. [7] use the stability of the output of regularized empirical risk minimization algorithms to show that by adding Laplace-distributed

noise to an empirical estimator θ or by adding an additional random term to the empirical risk $\frac{1}{n} \sum_{i=1}^n \ell(X_i, \theta)$, it is possible to obtain differential privacy and consistency of $\hat{\theta}$. Dwork and Lei [16] obtain similar results using robust statistical estimators, and Smith [40] shows that if one has suitably unbiased estimators, then differential privacy is possible without compromising asymptotic rates of convergence. Rubinstein et al. [38] use similar stability and perturbation techniques to demonstrate that it is possible to obtain differential privacy when solving support vector machine problems, and also show that if the desired privacy level α in the definition (2) is too small, it is actually impossible to obtain a parameter $\hat{\theta}_n$ minimizing the risk R .

Our goal is to understand the fundamental tradeoffs between maintaining privacy while still providing a useful output from the statistical learning procedure \mathcal{M} . Though intuitively there must be some tradeoff, quantifying it precisely has been difficult. As mentioned above, Rubinstein et al. [38] are able to show that it is impossible to obtain what they call an (ϵ, δ) -useful parameter vector θ that enjoys any differential privacy guarantees; however, it is unknown whether or not their guarantees might be improvable. Hall et al. [22] obtain minimax rates of convergence for differentially private histogram estimation, showing that if a histogram has d bins and we must guarantee α -differential privacy (2), then the minimax L^1 -risk of the histogram estimator is $d/(n\alpha)$, and Hardt and Talwar [23] give similar lower bounds on the amount of noise necessary to answer linear database queries. Blum et al. [5] also give lower bounds on the closeness of certain statistical quantities computed from the dataset, though their upper and lower bounds do not match. Sankar et al. [39] provide rate-distortion theorems for utility models involving information-theoretic quantities, which has some similarity to our risk-based framework, but it appears somewhat challenging to explicitly map their setting onto ours. With the goal of characterizing what it means to be both useful and private, Ghosh et al. [20] show that for a one-time computation of counts on a dataset X_1, \dots, X_n (i.e., the number of variables satisfying $X_i \in C$ for some set C), perturbing the output of a counting function using geometrically distributed noise is the unique optimal way to guarantee differential privacy while maximizing a natural notion of utility.

1.2 Our setting

In contrast to the above work, we study a more local notion of privacy [18, 27], in which each datum X_i is kept private from the method \mathcal{M} . The goal of many types of privacy is to guarantee that the output $\hat{\theta}_n$ of the method \mathcal{M} based on the data cannot be used to discover information about the individual samples X_1, \dots, X_n , but *locally private* algorithms only access disguised views of each datum X_i . Local algorithms are among the most classical approaches to privacy, tracing back to work on randomized response in the statistical literature [43], and rely on communication only of some disguised view Z_i of each true sample X_i .

Locally private algorithms are natural when the providers of the data—the population sampled to give X_1, \dots, X_n —do not even trust the statistician or statistical method \mathcal{M} , but the providers are interested in the parameter vector θ^* that minimizes the risk function. For example, in medical applications, a participant may be embarrassed about his use of drugs, or perhaps about his marital status, but if the loss ℓ is able to measure the likelihood of developing cancer, then the participant has high utility for access to the optimal parameters θ^* . Internet applications, where a user’s activity is logged across multiple websites or searches, provide another example: the user has a utility for a search engine to have a ranking function θ that returns relevant results for web searches, yet may not wish to reveal his or her search data. In essence, we would like the statistical procedure \mathcal{M} to learn *from* the data X_1, \dots, X_n but not *about* it.

The work most related to ours seems to be that of Kasiviswanathan et al. [27], who show that that (in some settings) locally private algorithms coincide with concepts that can be learned with polynomial sample complexity in Kearns’s statistical query (SQ) model [28]. This result is powerful, but has some limitations, as the statistical query model relies exclusively on count queries. In contrast, our analysis applies to estimators deriving from a broad class of convex risks (1) and it provides sharp rates of convergence.

We develop our approach to local privacy in the setting of two related privacy measures. The first is a worst-case measure of mutual information, where we view privacy preservation as a game between the providers of the data—who wish to preserve privacy—and nature. The second is based on differential privacy, where the provider of each datum communicates—subject to some constraints we make explicit later—the *most* differentially private view Z_i of his or her datum X_i .

Turning first to the information-theoretic formulation, and recalling that the method \mathcal{M} sees only the perturbed version Z_i of X_i , we use a uniform variant of mutual information $I(Z_i; X_i)$ between the random variables X_i and Z_i as our measure for privacy. Using mutual information and related information-theoretic ideas in the privacy and security context is by no means original; see, for example, the survey [31]. It is important to note, however, that standard mutual information has deficiencies as a measure of privacy [e.g., 18]. Accordingly, our uniform notion of mutual information is as follows: we say that the distribution Q generating Z from X is private only if $I(X; Z)$ is small for all possible distributions P on X , possibly subject to some constraints.

In this setting, we design procedures that allow consistent estimation of the parameter θ^* minimizing $R(\theta) = \mathbb{E}_P[\ell(X, \theta)]$, for any convex loss ℓ and distribution P on the data X . One central consequence of our analysis is a sharp characterization of the *excess risk*

$$\Delta_n(\hat{\theta}; \ell, \Theta) := \mathbb{E} \left[R(\hat{\theta}(Z_1, \dots, Z_n)) \right] - \inf_{\theta \in \Theta} R(\theta) \quad (3)$$

associated with any estimator $\hat{\theta}$ that satisfies a pre-specified privacy constraint. For particular collections \mathfrak{L} of loss functions $\ell \in \mathfrak{L}$, we bound the minimax convergence rate of all estimation procedures. More precisely, if ones wishes to guarantee a level of privacy $I(X_i; Z_i) \leq I^*$, then we show that there exists a constant $a(\mathfrak{L}, \Theta) \in \mathbb{R}_+$ —dependent only on the properties of the collection \mathfrak{L} and domain Θ —such that *for any* estimator $\hat{\theta}$ for the family \mathfrak{L} , the excess risk is lower bounded as

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\hat{\theta}; \ell, \Theta) \geq \frac{a(\mathfrak{L}, \Theta)}{\sqrt{nI^*}}. \quad (4a)$$

Moreover, we also prove that there exists another constant $b(\mathfrak{L}, \Theta) \geq a(\mathfrak{L}, \Theta)$ and provide explicit estimators $\hat{\theta}$ with privacy guarantee I^* such that

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\hat{\theta}; \ell, \Theta) \leq \frac{b(\mathfrak{L}, \Theta)}{\sqrt{nI^*}}. \quad (4b)$$

Turning to the setting of differential privacy, we are able to show similar results to the bounds (4a) and (4b). Namely, there exist constants $b'(\mathfrak{L}, \Theta) \geq a'(\mathfrak{L}, \Theta)$ such that if we wish to guarantee α -differential privacy, then for any estimator $\hat{\theta}$, the risk is lower bounded by

$$\sup_{\ell \in \mathfrak{L}} \Delta_n(\hat{\theta}; \ell, \Theta) \geq \frac{a'(\mathfrak{L}, \Theta)}{\alpha\sqrt{n}}, \quad (5a)$$

while there exist estimators $\hat{\theta}$ such that

$$\sup_{\ell \in \mathcal{L}} \Delta_n(\hat{\theta}; \ell, \Theta) \leq \frac{b'(\mathcal{L}, \Theta)}{\alpha \sqrt{n}}. \quad (5b)$$

Finally, we show that stochastic gradient descent is one procedure that achieves the above upper bounds, and moreover, that the ratios $b(\mathcal{L}, \Theta)/a(\mathcal{L}, \Theta)$ and $b'(\mathcal{L}, \Theta)/a'(\mathcal{L}, \Theta)$ are bounded above by a universal (numerical) constant. The bounds (4) and (5) thus establish and quantify explicitly the sharp tradeoff between learning and statistical estimation and the amount of privacy provided to the population. Moreover, the algorithms we use to give the upper bounds apply in streaming and online settings, requiring only a fixed-size memory footprint.

Our subsequent analysis will build on this favorable property of gradient-based methods. Indeed, in the remainder of the paper, we will assume that the communication protocol by which data is conveyed to the learner \mathcal{M} is based on (sub)gradients of the loss. As further motivation for this choice, note that the subgradient (more generally, a score function) of the loss ℓ is asymptotically sufficient in the sense of Le Cam [29]. A bit more precisely, gradients (in an asymptotic sense) contain *all* of the statistical information for risk minimization problems. Secondly, estimation procedures based on stochastic gradient information are asymptotically efficient [36], in the sense of both Bahadur and minimax efficiency [41, Chapter 8], and are thus essentially sample optimal; they also have minimax-optimality guarantees in finite-sample settings [1]. Moreover, many—perhaps most—estimation procedures are gradient-based [33, 6], and distributed optimization procedures that send subgradient information across a network to a centralized procedure \mathcal{M} are natural [e.g. 3]. Our arguments will also show that disguising subgradients is (in many settings) equivalent to disguising the data X itself.

1.3 Outline and techniques

We spend the remainder of the paper deriving the bounds displayed in (4) and (5). Our route to obtaining these bounds is based on a two-part analysis. First, we consider saddle points of the mutual information $I(X; Z)$, when viewed as a function of the distribution P of X and the conditional distribution $Q(\cdot | X)$ of Z , under natural constraints that still allow estimation. We consider related saddle points for differentially private conditional distributions. Having computed these saddle points, we can apply information-theoretic techniques for obtaining lower bounds on estimation [45, 1] to prove the results of the form (4b) or (5b). Our upper bounds then follow by application of known convergence rates for computationally efficient methods, such as the stochastic gradient and mirror descent algorithms [33, 34].

The remainder of the paper is organized as follows. We give a precise definition of our notions of local privacy in Section 2. Section 3 is devoted to information-theoretic lower bounds on the convergence rate of any statistical method \mathcal{M} in terms of the mutual information I^* between what the method \mathcal{M} observes and each sample X_i . We characterize the unique privacy guaranteeing distributions in Section 4, which provides a constructive mechanism for trading off privacy and learning. We devote Section 5 to the proofs of results given in Section 3, with our more technical results deferred to the appendices. We present our conclusions in Section 6.

Notation Before continuing, we give our notation and a few standard definitions. The *Kullback-Leibler (KL)* divergence between distributions P and Q defined on a set S , where P and Q are

assumed to have densities p and q with respect to a base measure ν^1 is given by

$$D_{\text{kl}}(P\|Q) := \int_S p(s) \log \frac{p(s)}{q(s)} d\nu(s).$$

Similarly, the *total-variation distance* between the distributions P and Q is defined as

$$\|P - Q\|_{\text{TV}} := \sup_{A \subset S} |P(A) - Q(A)| = \frac{1}{2} \int_S |p(s) - q(s)| d\nu(s).$$

For a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$, the subgradient set $\partial f(\theta)$ of f at the point θ is

$$\partial f(\theta) := \left\{ g \in \mathbb{R}^d : f(\theta') \geq f(\theta) + g^\top (\theta' - \theta), \text{ for all } \theta' \in \mathbb{R}^d \right\}.$$

We use $\partial \ell(x, \theta)$ to denote the subgradient set of the function $\theta \mapsto \ell(x, \theta)$, and for a convex function, $\nabla \ell(x, \theta)$ denotes an arbitrary element of $\partial \ell(x, \theta)$. We say that a function f is L -Lipschitz with respect to the norm $\|\cdot\|$ over the set Θ if

$$|f(\theta) - f(\theta')| \leq L \|\theta - \theta'\| \quad \text{for all } \theta, \theta' \in \Theta.$$

The notation $\|\cdot\|_p$ denotes a standard ℓ_p -norm. We use the abbreviation r.c.d. throughout for regular conditional distribution [4]. The extreme points of a set $C \subset \mathbb{R}^d$ are denoted by $\text{Ext}(C)$, the convex hull of C is denoted by $\text{Conv}(C)$, and the support of a distribution P is denoted $\text{supp } P$. We say values $a_n \asymp b_n$ if $\lim_n (a_n/b_n) = 1$. The symbol e_i denotes the i th standard basis vector in \mathbb{R}^d . Lastly, the symbol \rightrightarrows denotes a set-valued mapping [24].

2 Problem Formulation

We begin with a formal description of the communication protocol by which information about the random variables X is communicated to the procedure \mathcal{M} . We then define the notion of *optimal local privacy* studied in this paper, and the minimax framework in which we state our main results.

2.1 Communication protocol

In this paper, we focus on statistical learning procedures that have access to data through the subgradients $\partial \ell(X, \theta)$ of the loss functions. More formally, at each round, the method \mathcal{M} is given access to a random vector Z_i such that

$$\mathbb{E}[Z_i \mid X_i, \theta] \in \partial \ell(X_i, \theta), \tag{6}$$

where $\theta \in \Theta$ is a parameter chosen by the method. In Appendix A we present an argument that shows that the unbiasedness of the subgradient inclusion (6) is not only intuitively appealing but is, in a certain sense, necessary and cannot be avoided.

In detail, our communication protocol consists of the following steps:

- the method \mathcal{M} sends the parameter vector θ to the owner of the i th sample X_i ;

¹This is no loss of generality, as P and Q are absolutely continuous with respect to $\nu = \frac{1}{2}(P + Q)$.

- owner i computes a subgradient vector $g \in \partial\ell(X_i, \theta)$;
- the vector Z_i is communicated to \mathcal{M} under the constraint that $\mathbb{E}[Z_i \mid X_i, \theta] \in \partial\ell(X_i, \theta)$.

We assume throughout that there is a compact set $C \subset \mathbb{R}^d$ such that $\partial\ell(x, \theta) \subseteq C$ for all pairs $(\theta, x) \in \Theta \times \mathcal{X}$. Our goal is “disguise” the subgradient information with a random variable Z satisfying $Z \in D$, for some compact set D such that $C \subset \text{int } D \subset \mathbb{R}^d$.

For instance, a common choice of these sets are norm balls, say of the form

$$C = \{g \in \mathbb{R}^d : \|g\| \leq L\}, \quad \text{and} \quad D = \{g \in \mathbb{R}^d : \|g\| \leq M\},$$

where $\|\cdot\|$ is a given norm on \mathbb{R}^d , and the radius choice $M > L$ ensures that $C \subset \text{int } D$. This choice covers a variety of online optimization and stochastic approximation algorithms [49, 2, 34, 12, 1], for which it is assumed that for any $x \in \mathcal{X}$ and $\theta \in \Theta$, if $g \in \partial\ell(x, \theta)$ then $\|g\| \leq L$ for some norm $\|\cdot\|$. We may obtain privacy by allowing a perturbation of the subgradient g , which is then required to live in a (larger) norm ball of radius $M > L$.

2.2 Optimal local privacy

Suppose that X has distribution P , and for each $x \in \mathcal{X}$, let $Q(\cdot \mid x)$ denote the regular conditional probability measure of Z given that $X = x$. This pair defines the marginal distribution $Q(\cdot)$ via $Q(A) = \mathbb{E}[Q(A \mid X)]$, where the expectation taken with respect to $X \sim P$. The mutual information between X and Z is the expected Kullback-Leibler (KL) divergence between $Q(\cdot \mid X)$ and $Q(\cdot)$:

$$I(P, Q) = I(X; Z) := \mathbb{E}_P [D_{\text{kl}}(Q(\cdot \mid X) \| Q(\cdot))]. \quad (7)$$

We view the problem of privacy as a game between the adversary controlling P and the data owners, who use Q to obscure the samples X . In particular, we say a distribution Q guarantees a level of privacy I^* if and only if $\sup_P I(P, Q) \leq I^*$. Note that this guarantee is worst-case, ensuring that for any choice of distribution P , the publically available random variable Z provides at most mutual information I^* about the sample X .

Our goal is to find a saddle point P^*, Q^* such that

$$\sup_P I(P, Q^*) \leq I(P^*, Q^*) \leq \inf_Q I(P^*, Q), \quad (8)$$

where the first supremum is taken over all distributions P on X such that $\nabla\ell(X, \theta) \in C$ with P -probability 1, and the infimum is taken over all regular conditional distributions Q such that if $Z \sim Q(\cdot \mid X)$, then $Z \in D$ and $\mathbb{E}_Q[Z \mid X, \theta] = \nabla\ell(X, \theta)$. Indeed, if we can find P^* and Q^* satisfying the saddle point (8), then combination with the trivial direction of the max-min inequality yields

$$\sup_P \inf_Q I(P, Q) = I(P^*, Q^*) = \inf_Q \sup_P I(P, Q).$$

To fully formalize this idea and our notions of privacy, we define two collections of probability measures and associated losses. For sets $C \subset D \subset \mathbb{R}^d$, we define the source set

$$\mathcal{P}(C) := \{\text{Distributions } P \text{ such that } \text{supp } P \subset C\} \quad (9a)$$

and the set of regular conditional distributions (r.c.d.’s), or communicating distributions,

$$\mathcal{Q}(C, D) := \left\{ \text{r.c.d.’s } Q \text{ s.t. } \text{supp } Q(\cdot \mid c) \subset D \text{ and } \int_D z dQ(z \mid c) = c \text{ for } c \in C \right\}. \quad (9b)$$

The definitions (9a) and (9b) formally define the sets over which we may take infima and suprema in the saddle point calculations, and they capture what may be communicated. The conditional distributions $Q \in \mathcal{Q}(C, D)$ are defined so that for any loss ℓ with $\nabla \ell(x, \theta) \in C$, we have

$$\mathbb{E}_Q[Z \mid X, \theta] := \int_D z dQ(z \mid \nabla \ell(x, \theta)) = \nabla \ell(x, \theta).$$

We now make the following key definition:

Definition 1. The conditional distribution Q^* satisfies *optimal local privacy* for the sets $C \subset D \subset \mathbb{R}^d$ at level I^* if

$$\sup_P I(P, Q^*) = \inf_Q \sup_P I(P, Q) = I^*,$$

where the supremum is taken over distributions $P \in \mathcal{P}(C)$ and the infimum is taken over regular conditional distributions $Q \in \mathcal{Q}(C, D)$.

We also formulate a corresponding notion of local optimality in the differential privacy setting. For given sets $C \subset D$, define the differential privacy measure

$$\alpha^*(C, D) := \inf_Q \log \left[\sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q(S \mid X = x)}{Q(S \mid X = x')} \right], \quad (10)$$

where the infimum is taken over all regular conditional distributions $Q \in \mathcal{Q}(C, D)$ such that $\mathbb{E}_Q[Z \mid X = x] = x$. We define optimal local differential privacy as follows:

Definition 2. The conditional distribution Q^* satisfies *optimal local differential privacy* for the sets $C \subset D \subset \mathbb{R}^d$ if

$$\sup_P I(P, Q^*) = \inf_Q \sup_P I(P, Q),$$

where the supremum is taken over all distributions $P \in \mathcal{P}(C)$, and the infimum is taken over all $\alpha^*(C, D)$ -differentially private regular conditional distributions $Q \in \mathcal{Q}(C, D)$.

If a distribution Q^* satisfies optimal local privacy or optimal local differential privacy, then it guarantees that even for the worst possible distribution on X , the information communicated about X is limited. (Part of our results consist in showing that for suitable sets $C \subset D$, it is possible to attain $\alpha^*(C, D)$, so it is sensible to, in addition, choose the distribution that minimizes mutual information.)

In a sense, Definitions 1 and 2 capture the natural competition between privacy and learnability. The method \mathcal{M} specifies the set D to which the data Z it receives must belong; the “teachers,” or owners of the data X , choose the distribution Q to guarantee as much privacy as possible subject to this constraint. Using these mechanisms, if we can characterize a unique distribution Q^* attaining the infimum (8) for P^* (and by extension, for any P), then we may study the effects of requiring a bounded amount of information to be communicated to the method \mathcal{M} about X , which we do in Section 3.

2.3 Minimax error

Given an estimate $\hat{\theta}$ based on n samples X from a distribution P , we assess its quality in terms of the risk function $R(\theta) = \mathbb{E}[\ell(X, \theta)]$. In this section, we describe the minimax framework for obtaining bounds uniformly over all possible estimators.

More precisely, let \mathcal{M} denote any statistical procedure or method that operates on stochastic gradient samples, and let $\hat{\theta}_n$ denote the output of \mathcal{M} after receiving n such samples. The excess risk of the method \mathcal{M} on the risk $R(\theta)$ after receiving n sample gradients is given by

$$\epsilon_n(\mathcal{M}, \ell, \Theta, P) := R(\hat{\theta}_n) - \inf_{\theta \in \Theta} R(\theta) = \mathbb{E}_P[\ell(X, \hat{\theta}_n)] - \inf_{\theta \in \Theta} \mathbb{E}_P[\ell(X, \theta)]. \quad (11)$$

Note that this excess risk is a random variable, since the output $\hat{\theta}_n$ of the method is a random variable.

In our settings, in addition to the randomness in the sampling distribution P , there is additional randomness from the perturbation applied to stochastic gradients of the objective $\ell(X, \cdot)$ to mask X from the statistician or method \mathcal{M} . Let Q denote the regular conditional probability—the channel distribution—whose conditional part is defined on the range of the (set-valued) subgradient mapping $\partial\ell(X, \cdot) : \Theta \rightrightarrows \mathbb{R}^d$. Since the output $\hat{\theta}_n$ of the statistical procedure \mathcal{M} is a random function of both P and Q , we take the expectation and measure the expected sub-optimality of the risk according to P and Q . We let \mathfrak{L} denote a collection of loss functions, where for a distribution P on \mathcal{X} , the set $\mathfrak{L}(P)$ denotes the losses $\ell : \text{supp } P \times \Theta \rightarrow \mathbb{R}_+$ belonging to \mathfrak{L} . The *minimax error* is then given by

$$\epsilon_n^*(\mathfrak{L}, \Theta) := \sup_P \inf_{\mathcal{M}} \sup_{\ell \in \mathfrak{L}(P)} \mathbb{E}_{P, Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)], \quad (12)$$

where the expectation is taken over the random samples $X \sim P$ and $Z \sim Q(\cdot \mid X, \theta)$. In this paper, we provide characterizations of the minimax error (12) for several classes of loss functions $\mathfrak{L}(P)$, giving sharp results when the privacy distribution Q satisfies optimal local privacy for any loss function $\ell \in \mathfrak{L}(P)$ and distribution P .

3 Optimal Learning Rates and Tradeoffs

With our framework in place, we now turn to statements of our main results. We begin by imposing certain (weak) conditions on the families of loss functions that we consider, and subsequently turn to the main results of this section (Theorems 1, 2, and 3) as well as some of their consequences (Corollaries 1, 2, and 3).

3.1 Families of loss functions

We assume that our collection of loss functions obey certain natural smoothness conditions. For each $p \in [1, \infty]$, we use $\|\cdot\|_p$ to denote the usual ℓ_p -norm, and we use $q = \frac{p}{p-1}$ to denote the conjugate exponent satisfying the relation $1/p + 1/q = 1$. With this notation, we have the following definition:

Definition 3. For parameters $L > 0$ and $p \geq 1$, an (L, p) -loss function is a measurable function $\ell : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ such that for P -almost every $x \in \mathcal{X}$, the function $\theta \mapsto \ell(x, \theta)$ is convex and L -Lipschitz continuous with respect to the norm $\|\cdot\|_q$.

A convex loss ℓ satisfies Definition 3 if and only if for all $\theta \in \Theta$, we have the inequality $\|g\|_p \leq L$ for any subgradient $g \in \partial\ell(x, \theta)$ (e.g. [24]).

In order to illustrate this definition, let us consider a few examples:

Example 1. As a simple example, we may consider finding a multi-dimensional median, in which case the data $x \in \mathbb{R}^d$ and

$$\ell(x, \theta) = L \|\theta - x\|_1.$$

This loss is L -Lipschitz with respect to the ℓ_1 -norm, its subgradients belonging to $[-L, L]^d$, and hence ℓ belongs to the class of (L, ∞) -loss functions.

Example 2 (Classification). We may also consider classification based on either the hinge loss or logistic regression loss. In this setting, the data comes in pairs $x = (a, b)$, where $a \in \mathbb{R}^d$ is the set of regressors or predictors and $b \in \{-1, 1\}$ is the label; the losses are

$$\ell(x, \theta) = [1 - b \langle a, \theta \rangle]_+ \quad \text{and} \quad \ell(x, \theta) = \log(1 + \exp(-b \langle a, \theta \rangle)).$$

By computing (sub)gradients, we may verify that each of these belong to the class of (L, p) -losses if and only if the covariate vector $a \in \mathbb{R}^d$ satisfies $\|a\|_p \leq L$, which is a common assumption [7, 12, 38].

Definition 3 is natural given the communication strategy we outline in Section 2.1. Since our loss functions satisfy $\|\partial\ell(X, \theta)\| \leq L$, the channel distribution Q amounts to perturbing subgradients to larger norm balls while maintaining the appropriate expectations.

3.2 Bounds on minimax errors

We now state our three main theorems, deferring proofs to Section 5. Our first theorem applies to the class of (L, ∞) loss functions as given in Definition 3. For this theorem, we assume that the set to which the perturbed data Z must belong is $[-M_\infty, M_\infty]^d$, where $M_\infty \geq L$. In the notation of Definitions 1 and 2, this corresponds to taking $C = [-L, L]^d$ and $D = [-M_\infty, M_\infty]^d$. We state two variants of the first theorem, as one gives slightly sharper results for an important special case.

Theorem 1. *Let \mathfrak{L} be the collection of (L, ∞) loss functions, assume the conditions of the preceding paragraph, and let Q be optimally locally private (Definition 1) for \mathfrak{L} . Then*

(a) *If Θ contains the ℓ_∞ ball of radius r ,*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \frac{M_\infty r d}{\sqrt{n}}.$$

(b) *If $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$ and $d \geq 2$,*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{16} \frac{M_\infty r \sqrt{\log(2d)}}{\sqrt{n}}.$$

Our second main theorem applies to loss functions and objectives with a different geometry than the first and last. Now we assume that the loss functions \mathfrak{L} consist of $(L, 1)$ losses, and that the perturbed data must belong to the ℓ_1 ball of radius M_1 , i.e., $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M_1\}$. Thus

in the notation of Definition 1, we have $D = (M_1/L)C$, where $C = \{g \in \mathbb{R}^d : \|g\|_1 \leq L\}$. If we define $M = M_1/L$, we may define the constants

$$\gamma := \log \left(\frac{2d - 2 + \sqrt{(2d - 2)^2 + 4(M^2 - 1)}}{2(M - 1)} \right) \quad \text{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d - 1)}, \quad (13)$$

which are related to the unique distribution achieving optimal local privacy for the $(L, 1)$ losses and the larger ℓ_1 ball above (see equation (15) and Proposition 2). We have the following theorem.

Theorem 2. *Let \mathfrak{L} be the collection of $(L, 1)$ loss functions, assume the conditions of the preceding paragraph, and let Q be optimally private for the collection \mathfrak{L} . If Θ contains the ℓ_∞ -ball of radius r ,*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \frac{rL\sqrt{d}}{\sqrt{n}\Delta(\gamma)}.$$

For our final main theorem, we focus on differentially private algorithms, where we assume that communication respects optimal local differential privacy, as given by Definition 2. We use the same collection of loss functions \mathfrak{L} as in Theorem 1, that is, (L, ∞) -loss functions. We also assume that the set to which the perturbed data Z belong is $[-M_\infty, M_\infty]^d$, though the specific value of M_∞ is not actually important for the statement of the theorem.

Theorem 3. *Let \mathfrak{L} be the collection of (L, ∞) loss functions, and assume that Z is optimally locally differentially private (Definition 2), attaining α -differential privacy for the set \mathfrak{L} . Let $d \geq 2$ and assume $\alpha \leq 5/4$. Then*

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{\sqrt{d}}{\alpha} \frac{rL\sqrt{\log(2d)}}{32\sqrt{n}}.$$

Remarks: We make a few remarks on Theorems 1, 2, and 3. First, we note that, when reduced to the special case of having no random distribution Q , Theorems 1 and 2 each yield a minimax rate for stochastic optimization problems. Indeed, in Theorem 1, we may take $M_\infty = L$, in which case (focusing on the second statement of the theorem) we obtain that for $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{rL}{16} \sqrt{\frac{\log(2d)}{n}}.$$

Mirror descent algorithms [33, 34] can be used to minimize this class of loss functions, and their convergence rate matches this lower bound up to constant factors (also see our results in the sequel, as well as the explanation of Agarwal et al. [1]). Thus, our result when specialized to this setting is unimprovable. Moreover, our analysis is sharper than previous analyses, as none of the existing lower bounds recover the logarithmic dependence on the dimension d , which is evidently necessary.

In Theorem 2, if we take the constant $M_1 \downarrow L$, we see that $\gamma \rightarrow \infty$ and consequently $\Delta(\gamma) \rightarrow 1$. Thus we obtain that whenever Θ contains an ℓ_∞ ball of radius r ,

$$\epsilon_n^*(\mathfrak{L}, \Theta) \geq \frac{1}{163} \frac{rL\sqrt{d}}{\sqrt{n}}.$$

For this class of loss functions, the method of stochastic gradient descent attains a matching upper bound, again up to constant factors. (See Appendix C in Agarwal et al. [1], and also our results in the next section.)

Our second remark is that while our results appear to require disguising only gradient information, based on our communication setting in Section 2.1, this restriction is not actually substantial. Indeed, when the domain Θ is a norm ball we can establish each of our lower bounds using the loss function $\ell(x, \theta) = \langle x, \theta \rangle$. In this case, $\nabla \ell(x, \theta) = x$, so that the communication scheme explicitly disguises *exactly* the individual data X_i .

Finally, we have presented results for specific geometric properties of the loss functions \mathfrak{L} . These geometric properties are natural, as exemplified by our examples in Section 3.1. It is, however, also possible to use our techniques to derive alternative results; such extension requires computing the optimal distribution attaining local privacy according to Definitions 1 or 2, then applying the lower-bounding techniques we develop in the sequel.

3.3 Tradeoffs between privacy and statistical error

We now turn to some consequences of Theorems 1, 2, and 3 for the tradeoffs between rates of convergence for any statistical procedure and the desired privacy of a user. We present three corollaries that characterize this tradeoff. Looking ahead to Section 4, we may use Propositions 1, and 2, and 3 to establish the results. For the mutual-information-based results (Theorems 1 and 2), we apply the first two propositions to derive a bijection between the sizes M_∞ and M_1 of the perturbation set and the amount of privacy, as measured by the worst case mutual information I^* . We can then combine the lower bounds of Theorems 1 and 2 with results on stochastic approximation to obtain the tradeoffs. For differentially-private algorithms—whose lower bound is provided by Theorem 3—we can show an upper bound on the necessary magnitude of the gradient bound M_∞ to allow α -differential privacy, again applying known stochastic approximation results. We provide the full proofs in Sections 5.7, 5.8, and 5.9, respectively.

In each of our corollaries, the upper bound is attained by (a variant of) mirror descent [33, 2, 34], which is a non-Euclidean generalization of the stochastic gradient method [33, 36, 49]. Recall that stochastic gradient methods are iterative methods that update a parameter θ^t over iterations t of an algorithm using stochastic gradient information. In particular, at iteration t , the algorithm receives a vector $g_t \in \mathbb{R}^d$ with conditional expectation $\mathbb{E}[g_t \mid \theta^t] \in \partial R(\theta^t)$, then performs the update

$$\theta^{t+1} = \operatorname{argmin}_{\theta \in \Theta} \{ \eta \langle g_t, \theta \rangle + \Psi(\theta, \theta^t) \}.$$

Here η is a step-size and Ψ is a Bregman divergence, which keeps θ^{t+1} relatively close to θ^t . (See the papers [2, 34] for further details.) With appropriate choice of Ψ , the mirror descent algorithm enjoys the following convergence guarantees. Define $\hat{\theta}_n = \frac{1}{n} \sum_{t=1}^n \theta^t$. If $\mathbb{E}[\|g_t\|_\infty^2 \mid \theta^t] \leq M_\infty^2$ for all t and Θ is contained in the ℓ_1 -ball of radius r_1 , then with appropriate choice of Ψ and η

$$\mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) = \mathcal{O} \left(\frac{M_\infty r_1 \sqrt{\log d}}{\sqrt{n}} \right). \quad (14a)$$

See, for example, Beck and Teboulle [2, Section 5] or Nemirovski et al. [34, Section 2.3]. Similarly, with the choice $\Psi(\theta, \theta') = \|\theta - \theta'\|_2^2$, if $\mathbb{E}[\|g_t\|_2^2 \mid \theta^t] \leq M_2^2$ and Θ is contained in the ℓ_2 -ball of radius r_2 , then

$$\mathbb{E}[R(\hat{\theta}_n)] - R(\theta^*) = \mathcal{O} \left(\frac{M_2 r_2}{\sqrt{n}} \right). \quad (14b)$$

For instance, see the references [49, 34, Section 2.2] for results of this type.

Using the mirror descent algorithm, we can establish the following:

Corollary 1. *Under the conditions of Theorem 1(b), assume moreover that $M_\infty \geq 2L$ and Q^* satisfies optimal local privacy at information level I^* . Then for universal constants $0 < c_l \leq c_u$, the minimax error is sandwiched as*

$$c_l \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{\log d}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{\log d}}{\sqrt{n}}.$$

It is worth noting that similar upper and lower bounds can be obtained under the conditions of part (a) of Theorem 1, again by using mirror descent, but we lose a factor of $\sqrt{\log d}$ in the lower bound. (There is an additional factor of d in the statement (a), and $\Theta \supseteq \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq r/d\}$.) In this case we would not need to assume that Θ is an ℓ_1 -ball for the lower bound.

We now turn to an analogous result, but based on an application of Theorem 2 and Proposition 2.

Corollary 2. *Under the conditions of Theorem 2, assume that $M_1 \geq 2L$ and Q^* satisfies optimal local privacy at information level I^* . Moreover, suppose that Θ contains an ℓ_∞ -ball of radius $c_1 r$ and is contained in an ℓ_∞ -ball of radius $c_2 r$, where $0 < c_1 \leq c_2$ are constants. Then for universal constants $0 < c_l \leq c_u$, the minimax error is sandwiched as*

$$c_l \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{d}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\sqrt{I^*}} \cdot \frac{rL\sqrt{d}}{\sqrt{n}}.$$

Finally, we provide a corollary to Theorem 3, which gives us sharp tradeoffs in the differentially private case.

Corollary 3. *Under the conditions of Theorem 3, assume that Q^* satisfies Definition 2, attaining α -differential privacy. Then for universal constants $0 < c_l \leq c_u$, the minimax error is sandwiched as*

$$c_l \frac{\sqrt{d}}{\alpha} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}} \leq \epsilon_n^*(\mathfrak{L}, \Theta) \leq c_u \frac{\sqrt{d}}{\alpha} \cdot \frac{rL\sqrt{\log(2d)}}{\sqrt{n}}.$$

As noted above, mirror descent (or stochastic gradient descent) achieves the upper bound in each of the corollaries. We also note the difference in the dimension dependence in the convergence rates given by Corollaries 1 and 3 and that given by Corollary 2. In particular, the former two have a dependence on dimension growing as \sqrt{d} , while the latter depends on d . This is somewhat intuitive: under the conditions of Theorems 1(b) and 3, we are in a high-dimensional regime with a small set Θ (see, e.g. [2, Section 5], [33, Chapter 5], or [34, Section 2.3]). So we expect weaker dimension dependence. In Corollary 2, any optimization method must essentially identify d different coordinates of a vector in $[-r, r]^d$, an ℓ_∞ ball, which causes slowness, and yields a scaling of $\sqrt{d}rL/\sqrt{n}$ even in the standard (non-private) minimax case [1]. Thus for both Corollaries 1 and 2 we see that incorporating privacy induces a penalty of roughly $\sqrt{d}/\sqrt{I^*}$ in convergence rate. The scaling differences for mutual information (as $1/\sqrt{I^*}$) and differential privacy (as $1/\alpha$) are—as yet—incomparable, as there does not appear to be a simple mapping between information-theoretic notions of privacy and differential privacy.

4 Saddle Points, Optimal Privacy, and Mutual Information

In this section, we explore conditions for a distribution Q^* to satisfy optimal local privacy, as given by Definition 1. We give a few characterizations of necessary (and sometimes sufficient) conditions based on the compact sets $C \subset D$ for distributions P^* and Q^* to achieve the saddle point (8). Our results can be viewed as rate distortion theorems [21, 8, 10] (with source P and channel Q) for certain compact alphabets, though as far as we know, they are all new. Thus, we refer to the conditional distribution Q , which is designed to maintain the privacy of the data X by communication of Z , interchangeably as the privacy-preserving distribution or the channel distribution.

Note that since we wish to bound $I(X; Z)$ for general losses ℓ , as captured in the definitions of the source $\mathcal{P}(C)$ and communication set $\mathcal{Q}(C, D)$ in Eqs. (9a) and (9b), we must address the case when $\ell(X, \theta) = \langle \theta, X \rangle$, in which case $\nabla \ell(X, \theta) = X$; this shows (by the data-processing inequality [21, Chapter 5]) that it is no loss of generality to assume that $X \in C$ with probability 1 and that we must have $\mathbb{E}[Z | X] = X$. Thus we present each of our results assuming that $\ell(X, \theta) = \langle \theta, X \rangle$, since a distribution Q^* is optimally locally private or optimally differentially locally private if and only if it attains the saddle point with *this* choice of loss.

4.1 General saddle point characterizations

We begin with a general characterization, first defining the types of sets C and D that we use in our characterization of privacy. Such sets are reasonable for many applications (recall Section 3.1). We focus on the case when the compact sets C and D are (suitably symmetric) norm balls:

Definition 4. Let $C \subset \mathbb{R}^d$ be a compact convex set with extreme points $u_i \in \mathbb{R}^d$, $i \in I$ for some index set I . Then C is a *rotationally invariant through its extreme points* if $\|u_i\|_2 = \|u_j\|_2$ for each i, j , and for any unitary matrix U such that $Uu_i = u_j$ for some $i \neq j$, then $UC = C$.

Some examples of convex sets rotationally invariant through their extreme points include ℓ_p -norm balls for $p = 1, 2, \infty$, though ℓ_p -balls for $p \notin \{1, 2, \infty\}$ are not.

The following theorem gives a general characterization of the minimax mutual information for such rotationally invariant sets by providing saddle point distributions P^* and Q^* . We provide the proof of Theorem 4 in Section D.1.

Theorem 4. Let C be a compact convex polytope rotationally invariant through its $m < \infty$ extreme points $\{u_i\}_{i=1}^m$ and $D = (1 + \kappa)C$ for some $\kappa > 0$. Let Q^* be the conditional distribution of $Z | X$ that maximizes the entropy $H(Z | X = x)$ subject to the constraints that

$$\mathbb{E}_Q[Z | X = x] = x$$

for $x \in C$ and that Z is supported on $(1 + \alpha)u_i$ for $i = 1, \dots, m$. Then Q^* satisfies Definition 1, optimal local privacy, and Q^* is (up to measure zero sets) unique. Moreover, the distribution P^* uniform on $\{u_i\}_{i=1}^m$ attains the saddle point (8).

Remarks: We make a few brief remarks here, deferring a somewhat deeper discussion of the implications of Theorem 4 to Section D.1, as an understanding of the proof helps. While in the theorem we assume that $Q^*(\cdot | X = x)$ maximizes the entropy for each $x \in C$, this is not in fact essential. In fact, we may introduce a X' between X and Z : let X' be distributed among the

extreme points $\{u_i\}_{i=1}^m$ of C in *any* way such that $\mathbb{E}[X' | X] = X$, then use the maximum entropy distribution $Q^*(\cdot | u_i)$ defined in the theorem when $X \in \{u_i\}_{i=1}^m$ to sample Z from X' . By using the convexity of the negative entropy as in the bound (44) in the proof of Theorem 4 (really, the information processing inequality [21, Chapter 5]), this Markov chain $X \rightarrow X' \rightarrow Z$ guarantees at least minimax information $I(X; Z) \leq \inf_Q \sup_P I(P, Q)$.

4.2 Specific saddle point computations

With Theorem 4 in place, we can explicitly characterize the minimax mutual information for ℓ_1 and ℓ_∞ balls by computing maximum entropy distributions. That is, we show the unique distributions that attain optimal local privacy—the distributions that guarantee as much (of our definition of) privacy as possible subject to certain constraints. We present two propositions in this regard, providing some discussion and giving proofs in Sections D.2 and D.3.

First, consider the case where $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$. For notational convenience, we define the binary entropy $h(p) = -p \log p - (1-p) \log(1-p)$. We have

Proposition 1. *For a constant $M \geq 1$, let $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$ be random variables such that $\mathbb{E}[Z | X] = X$ almost surely. Define Q^* to be the conditional distribution on $Z | X$ such that the coordinates of Z are independent, have range $\{-M, M\}$, and*

$$Q^*(Z_i = M | X) = \frac{1}{2} + \frac{X_i}{2M} \quad \text{and} \quad Q^*(Z_i = -M | X) = \frac{1}{2} - \frac{X_i}{2M}.$$

Then Q^* satisfies Definition 1, optimal local privacy, and moreover,

$$\sup_P I(P, Q^*) = d - d \cdot h\left(\frac{1}{2} + \frac{1}{2M}\right).$$

Before continuing, we give a slightly more intuitive understanding of Proposition 1. Concavity implies that for $a, b > 0$, $\log(a) \leq \log b + b^{-1}(a - b)$, or $-\log(a) \geq -\log(b) + b^{-1}(b - a)$, so

$$-\log\left(\frac{1}{2} - \frac{1}{2M}\right) \geq -\log \frac{1}{2} + 2 \cdot \frac{1}{2M} \quad \text{and} \quad -\log\left(\frac{1}{2} + \frac{1}{2M}\right) \geq -\log \frac{1}{2} - 2 \cdot \frac{1}{2M}.$$

In particular, we see that

$$h\left(\frac{1}{2} + \frac{1}{2M}\right) \geq -\left(\frac{1}{2} + \frac{1}{2M}\right) \left(-\log 2 - \frac{1}{M}\right) - \left(\frac{1}{2} - \frac{1}{2M}\right) \left(-\log 2 + \frac{1}{M}\right) = \log 2 - \frac{1}{M^2}.$$

That is, we have for *any* distribution P on X , where $X \in [-1, 1]^d$, that (in natural logarithms)

$$I(P, Q^*) \leq \frac{d}{M^2},$$

and this bound is tight to $\mathcal{O}(M^{-3})$.

We now consider the case when $X \in \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$ and $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M\}$. Here the arguments are slightly more complicated, as the coordinates of the random variables are no longer independent, but Theorem 4 still allows us to explicitly characterize the saddle point of the mutual information. Before stating the proposition, we recall that if $e_i \in \mathbb{R}^d$ are the standard basis vectors, then the extreme points of the ℓ_1 -ball of radius 1 are the $2d$ vectors $\{\pm e_i\}_{i=1}^d$.

Proposition 2. For a constant $M > 1$, let $X \in \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$ and $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M\}$ be random variables. Define the parameter

$$\gamma := \log \left(\frac{2d - 2 + \sqrt{(2d - 2)^2 + 4(M^2 - 1)}}{2(M - 1)} \right), \quad (15)$$

and let Q^* be the conditional distribution on $Z \mid X$ such that Z is supported on $\{\pm Me_i\}_{i=1}^d$, and

$$Q^*(Z = Me_i \mid X = e_i) = \frac{e^\gamma}{e^\gamma + e^{-\gamma} + (2d - 2)}, \quad (16a)$$

$$Q^*(Z = -Me_i \mid X = e_i) = \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + (2d - 2)}, \quad (16b)$$

$$Q^*(Z = \pm Me_j \mid X = e_i, j \neq i) = \frac{1}{e^\gamma + e^{-\gamma} + (2d - 2)}. \quad (16c)$$

(For $X \notin \{\pm e_i\}$, define X' to be randomly selected in any way from among $\{\pm e_i\}$ such that $\mathbb{E}[X' \mid X] = X$, then sample Z from X' according to (16a)–(16c).) Then Q^* satisfies Definition 1, optimal local privacy, and

$$\sup_P I(P, Q^*) = \log(2d) - \log(e^\gamma + e^{-\gamma} + 2d - 2) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2}.$$

Proposition 2 is somewhat more complex than the ℓ_∞ case. We remark that the additional sampling to guarantee that $X' \in \{\pm e_i\}$ (where the conditional distribution Q^* is defined) can be accomplished simply: define the random variable X' so that $X' = e_i \text{sign}(x_i)$ with probability $|x_i|/\|x\|_1$. Evidently $\mathbb{E}[X' \mid X] = x$, and $X \rightarrow X' \rightarrow Z$ for Z distributed according to Q^* defines a Markov chain as in our remarks following Theorem 4. An asymptotic expansion allows us to gain a somewhat clearer picture of the values of the mutual information, though we do not derive upper bounds as we did for Proposition 1. We have the following corollary, proved in Appendix E.

Corollary 4. Let Q^* denote the conditional distribution in Proposition 2. Then

$$\sup_P I(P, Q^*) = \frac{d}{2M^2} + \Theta \left(\min \left\{ \frac{d^3}{M^4}, \frac{\log^4(d)}{d} \right\} \right).$$

4.3 Saddle points for differentially private communication

Our final result in this section characterizes saddle points for distributions satisfying Definition 2. Such calculations are, in general, non-trivial, so we restrict our attention to results necessary for the setting of Theorem 3. To that end, we focus on the case where C and D are ℓ_∞ balls, which is relevant for high-dimensional statistical and optimization settings. Without loss of generality (by scaling), we may take $C = [-1, 1]^d$ and $D = [-M, M]^d$. We have

Proposition 3. For a constant $M \geq 1$, let $X \in [-1, 1]^d$ and $Z \in [-M, M]^d$ be random variables such that $\mathbb{E}[Z \mid X] = X$ almost surely. Fix any $x \in \{-1, 1\}^d$ and for $k = \{0, 2, 4, \dots, 2 \lceil d/2 \rceil - 2\}$ define the constants $q_k^+ \geq 0$ and $q_k^- \geq 0$ to satisfy the linear equations

$$Mq_k^+ \sum_{z \in \{-1, 1\}^d : \langle z, x \rangle > k} z + Mq_k^- \sum_{z \in \{-1, 1\}^d : \langle z, x \rangle \leq k} z = x \quad \text{and} \quad q_k^+ + q_k^- = 1.$$

Set $k^* = \operatorname{argmin}_k \{q_k^+/q_k^-\}$. Define Q^* to be the distribution supported on $\{-M, M\}^d$ with probability mass function defined by

$$Q^*(Z = Mz \mid X = x) = \begin{cases} q_{k^*}^+ & \text{if } \langle z, x \rangle > k^* \\ q_{k^*}^- & \text{if } \langle z, x \rangle \leq k^* \end{cases} \quad (17)$$

for $z, x \in \{-1, 1\}^d$. (For $X \notin \{-1, 1\}^d$, define X' to be randomly chosen from $\{-1, 1\}^d$ such that $\mathbb{E}[X \mid X'] = X$, then sample Z according to the above p.m.f.)

If k^* is unique, then Q^* uniquely satisfies Definition 2, optimal local differential privacy. If k^* is non-unique, any distribution satisfying optimal local differential privacy is in the set of all convex combinations of distributions Q^* defined via (17) for k minimizing q_k^+/q_k^- .

The proof of Proposition 3 is technical, and we defer it to Section D.4. We make a few remarks, however. First, we provide a simplified explanation of the the linear equations in the proposition. By symmetry, no matter the value of $x \in \{-1, 1\}^d$ chosen, the same q_k^+ and q_k^- solve the linear equations. Proposition 3 shows the *structure* of the distribution attaining optimal local differential privacy. That is, the proposition shows that the distribution $Q^*(\cdot \mid x)$ assigns mass only on the points $z \in \{-M, M\}^d$, and moreover, it assigns one of two masses: either q^+ or q^- . Whether a point $z \in \{-M, M\}^d$ is assigned the higher or lower mass depends on its agreement with the initial point x being perturbed, that is, whether $\langle z, x \rangle > k/M$ or $\langle z, x \rangle \leq k/M$.

Thus, for a fixed level k , the amount of differential privacy α attained is given by $e^\alpha = q_k^+/q_k^-$, so that to find the most differentially private distribution, we calculate the minimizing k for q_k^+/q_k^- . Note also that q_k^+ is a non-decreasing function of k , and that while the minimizing k may be non-unique, if $a/b = c/d$, then we have for any $\lambda \in [0, 1]$ that $a/c = b/d$, so

$$\frac{\lambda a + (1 - \lambda)c}{\lambda b + (1 - \lambda)d} = \frac{\lambda bc/d + (1 - \lambda)c}{\lambda ad/c + (1 - \lambda)d} = \frac{c(\lambda b/d + (1 - \lambda))}{d(\lambda a/c + (1 - \lambda))} = \frac{c}{d} = \frac{a}{b}.$$

In particular, the convex combination of α -differentially private distributions from Proposition 3 is precisely α -differentially private. So Proposition 3 gives a mechanical way to compute the possible set of distributions satisfying optimal local differential privacy.

5 Proofs of Statistical Rates

In this section, we prove Theorems 1, 2, and 3 as well as Corollaries 1, 2, and 3. Our proofs build on classical information-theoretic techniques from statistical minimax theory [45, 46] as well as some intermediate results due to Agarwal et al. [1]. At a high level, our approach is as follows. Beginning with an appropriately chosen finite set \mathcal{V} , we assign a risk functions R_v to each member $v \in \mathcal{V}$. The resulting collection $\{R_v\}_{v \in \mathcal{V}}$ of risk functions is chosen so that they “separate” points in the set \mathcal{V} , meaning that if $\theta \in \Theta$ is a point that approximately minimizes the function R_v , then for any $w \neq v$, the point θ *cannot* also be an approximate minimizer of R_w . This separation property allows us to deduce that statistical estimation implies the existence of a testing procedure that distinguishes v for $w \neq v$. We then use Fano’s inequality to obtain a lower bound on the testing error, so that the final step is to obtain good upper bounds on the mutual information between the random variable X_i and the vector Z_i communicated.

5.1 Reduction to testing

We begin by describing the reduction from bounding the minimax error to a testing problem. It assumes a given collection of risk functions $\{R_v\}_{v \in \mathcal{V}}$ indexed by a finite set \mathcal{V} ; see Section 5.2 to follow for discussion of the particular collections used in our analysis. For each $v \in \mathcal{V}$, we choose some representative $\theta_v^* \in \underset{\theta \in \Theta}{\operatorname{argmin}} R_v(\theta)$ of the set of all minimizing vectors. Our reduction is based on a discrepancy measure between pairs of risk functions, first introduced by Agarwal et al. [1], defined as

$$\rho(R_v, R_w) := \inf_{\theta \in \Theta} [R_v(\theta) + R_w(\theta) - R_v(\theta_v^*) - R_w(\theta_w^*)], \quad (18)$$

The ρ -separation of the set \mathcal{V} is given

$$\rho^*(\mathcal{V}) := \min \{\rho(R_v, R_w) : v, w \in \mathcal{V}, v \neq w\}. \quad (19)$$

When the set \mathcal{V} is clear from context, we use ρ^* as shorthand for this separation. The key to the definition (19) is that the separation allows us to lower bound the expected optimality gap of a statistical method \mathcal{M} by the probability of error in a hypothesis test. First, note that for any $\theta \in \Theta$, there is at most one $v \in \mathcal{V}$ such that $R_v(\theta) - R_v(\theta_v^*) < \rho^*/2$. Indeed, if this inequality holds for both v and $w \neq v$,

$$\rho^*(\mathcal{V}) \leq R_v(\theta) + R_w(\theta) - R_v(\theta_v^*) - R_w(\theta_w^*) < \rho^*(\mathcal{V}),$$

a contradiction. The following result is a variant of Lemma 2 from Agarwal et al. [1]:

Lemma 1. *Let P be a joint distribution over $X \in \mathbb{R}^d$ and $V \in \mathcal{V}$ such that X are i.i.d. given V and*

$$\mathbb{E}_P[\ell(X, \theta) \mid V = v] = R_v(\theta).$$

Let Q be the conditional distribution of Z given the subgradients $\partial\ell(X, \cdot)$. For any minimization procedure \mathcal{M} , one may construct a hypothesis test $\hat{v}(\mathcal{M}) : (Z_1, \dots, Z_n) \rightarrow \mathcal{V}$ such that

$$\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{\rho^*(\mathcal{V})}{2} \mathbb{P}_{P,Q}[\hat{v} \neq V].$$

In particular, if we can bound the probability of error of any hypothesis test for identifying V based on stochastic subgradient samples Z_1, \dots, Z_n , then we have lower bounded the rate at which it is possible to minimize the risk R .

In order to prove a lower bound on the error of a hypothesis testing problem, we apply Fano's inequality [8]. Let $V \in \mathcal{V}$ be chosen uniformly at random from \mathcal{V} . If a procedure observes random variables Z_1, \dots, Z_n , Fano's inequality ensures that for any estimate \hat{v} of V —that is, any measurable function \hat{v} of Z_1, \dots, Z_n —the test error probability satisfies the lower bound

$$\mathbb{P}(\hat{v}(Z_1, \dots, Z_n) \neq V) \geq 1 - \frac{I(Z_1, \dots, Z_n; V) + \log 2}{\log |\mathcal{V}|}. \quad (20)$$

Using the lower bound provided by Lemma 1 and Fano's inequality (20), the structure of our remaining proofs becomes more apparent. Each lower bound argument proceeds in three steps:

- (1) We construct a collection of loss functions satisfying Definition 3, computing the minimal separation (19) so that we may apply Lemma 1. (See Sections 5.2.1–5.2.3.)

- (2) To be able to apply Fano's inequality (20), we provide an upper bound on the mutual information $I(Z_1, \dots, Z_n; V)$ for our specific choice of loss from step 1. To do so, we use the fact that for each of Theorems 1, 2, and 3, we used a distribution Q that satisfies our Definition 1 of optimal local privacy; this requires some subtlety in providing the bound. (See Lemmas 5, 6, 7, and 8 in Section 5.3.)
- (3) The final step is to use the results of steps 1 and 2 in the application of Lemma 1 and Fano's inequality (20). This then yields the theorems.

We provide the formal proofs of Theorems 1, 2, and 3 in Sections 5.4, 5.5, and 5.6 respectively; the next two sections are devoted to steps 1 and 2.

5.2 Collections of loss functions

In this section, we construct three example sets of functions, each yielding a different collection of risks, enumerating their separation properties to be able to apply Lemma 1.

5.2.1 Linear Losses

Our first collection of risk functionals is relatively simple, based on families of linear loss function. Θ . Assuming that the random variables X take values in \mathbb{R}^d , we define the linear loss functions

$$\ell(X, \theta) := \langle X, \theta \rangle = \sum_{j=1}^d X_j \theta_j. \quad (21)$$

For this collection of loss functions, we let $\mathcal{V} = \{\pm e_i\}_{i=1}^d$, where the vectors e_i are the standard basis vectors in \mathbb{R}^d , whence $|\mathcal{V}| = 2d$. We also fix a $\delta \in (0, 1/4]$, which we specify later, and choose the distribution P on X so that the final risk is equal to

$$R_v(\theta) = \mathbb{E}_P[\langle \theta, X \rangle] = \frac{c\delta}{d} \langle v, \theta \rangle. \quad (22)$$

We choose the constant c so that the linear loss functions (22) belong to the appropriate loss class.

To construct a risk of the form (22), we draw the random vector $X \in \mathbb{R}^d$ conditional on the parameter v , choosing X from among the 2^d vectors in the scaled hypercube $\{-c, c\}^d$ —viz.

$$\text{Choose } X \in \{-c, c\}^d \text{ with independent coordinates, where } X_j = \begin{cases} c/d & \text{w.p. } \frac{1+\delta v_j}{2} \\ -c/d & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases} \quad (23)$$

Under the sampling strategy (23), when $v = \pm e_i$, the coordinate X_j is independent and uniformly chosen from $\{-c/d, c/d\}$ for $j \neq i$. Additionally, we have that $\mathbb{E}[\ell(X, \theta)] = R_v(\theta)$, and moreover:

Lemma 2. *In the sampling scheme (23), with $c = Ld$:*

- (a) *The loss (21) is L -Lipschitz with respect to the ∞ -norm.*
- (b) *For the optimization domain $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, the ρ -separation of the set $\mathcal{V} = \{\pm e_i\}_{i=1}^d$ is $\rho^*(\mathcal{V}) = Lr\delta$.*

Proof The first statement of the lemma is immediate. For the second, we compute minimizers. Indeed, by definition of the dual norm, we see that for $v \in \mathcal{V}$,

$$\inf_{\|\theta\|_1 \leq r} R_v(\theta) = \inf_{\|\theta\|_1 \leq r} \frac{c\delta}{d} \langle v, \theta \rangle = -\frac{c\delta}{d} r \|v\|_\infty = -L\delta r,$$

and the minimizer is uniquely attained at $\theta_v^* = -rv$. Then we have for any $w \neq v$ that

$$\inf_{\|\theta\|_1 \leq 1} [\langle v + w, \theta \rangle] + \|v\|_\infty + \|w\|_\infty = -\|v + w\|_\infty + \|v\|_\infty + \|w\|_\infty \geq -1 + 1 + 1 = 1,$$

since no identical coordinates of v and w have the same sign. Multiplying the result by $Lr\delta$ completes the proof. \square

5.2.2 Hinge (SVM) Losses

We now turn to families of losses that are useful for analyzing the case of stochastic subgradients bounded in ℓ_1 -norm. Let $\mathcal{V} \subset \{-1, 1\}^d$ be a subset of the binary hypercube such that for all distinct pairs $v \neq v'$, we have $\|v - v'\|_1 \geq d/2$, or equivalently $\|v - v'\|_0 \geq d/4$. From the Gilbert-Varshamov bound (e.g. [46, Lemma 4]) there are sets of this form with cardinality at least $\text{card}(\mathcal{V}) \geq \exp(d/8)$.

For a fixed constant $c > 0$, we define the hinge loss

$$\ell(x, \theta) = c[r - \langle x, \theta \rangle]_+. \quad (24)$$

As our sampling process for the data, we choose X from among the $2d$ positive and negative standard basis vectors $\pm e_j$ —namely

$$\text{Choose index } j \in \{1, \dots, d\} \text{ uniformly at random, and set } X = \begin{cases} e_j & \text{w.p. } \frac{1+\delta v_j}{2} \\ -e_j & \text{w.p. } \frac{1-\delta v_j}{2}, \end{cases} \quad (25)$$

where $\delta \in (0, 1/4]$ is fixed. The combination of hinge loss (24) and sampling strategy (25) yields the risk function

$$R_v(\theta) := \frac{c}{d} \sum_{j=1}^d \frac{1+\delta v_j}{2} [r - \langle e_j, \theta \rangle]_+ + \frac{c}{d} \sum_{j=1}^d \frac{1-\delta v_j}{2} [r + \langle e_j, \theta \rangle]_+. \quad (26)$$

Assuming that Θ contains the ℓ_∞ ball of radius r , the (unique) minimizer of the risk over Θ is

$$\theta_v^* := \argmin_{\theta \in \Theta} R_v(\theta) = rv \in r\{-1, 1\}^d \subset \Theta.$$

Moreover, this risk has the following properties:

Lemma 3. *For any set $\Theta \supseteq [-r, r]^d$, we have:*

- (a) *For P with support $\text{supp } P \subseteq \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\}$, the loss function $\ell(x, \theta) = c[r - \langle \theta, x \rangle]_+$ is c -Lipschitz with respect to the ℓ_1 -norm.*
- (b) *If $v, w \in \mathcal{V}$ with $v \neq w$, the discrepancy $\rho(R_v, R_w) \geq rc\delta/2$.*

Proof The first claim is immediate (e.g. [24]), since $\|\partial\ell(x, \theta)\|_1 \leq c\|x\|_1 \leq c$. For the second statement of the lemma, we see that the minimum of

$$R_v(\theta) + R_w(\theta) = \frac{c}{d} \sum_{j=1}^d ([r - \langle e_j, \theta \rangle]_+ + [r + \langle e_j, \theta \rangle]_+) + \frac{c\delta}{d} \sum_{j:v_j=w_j} (v_j[r - \langle e_j, \theta \rangle]_+ - v_j[r + \langle e_j, \theta \rangle]_+)$$

is attained by any $\theta \in \mathbb{R}^d$ with $\theta_j \in [-r, r]$ for j such that $v_j \neq w_j$ and $\theta_j = rv_j$ for j such that $v_j = w_j$. Thus we have

$$\begin{aligned} \inf_{\theta \in \Theta} \{R_v(\theta) + R_w(\theta)\} - R_v(\theta_v^*) - R_w(\theta_w^*) &= \frac{c}{d} \sum_{j=1}^d 2r - \frac{2c}{d} \sum_{j:v_j=w_j} r\delta - cr(1-\delta) - cr(1-\delta) \\ &= 2cr - 2cr + 2cr\delta - \frac{2cr\delta}{d} (d - \|v - w\|_0) = \frac{2cr\delta}{d} \|v - w\|_0. \end{aligned}$$

Since $\|v - w\|_0 \geq d/4$ by construction, we have $\rho(R_v, R_w) \geq rc\delta/2$, as desired. \square

5.2.3 Median-type Losses

We now describe a class of median-type losses, one with more general applicability than the linear losses of Section 5.2.1. As in Section 5.2.2, let $\mathcal{V} \subset \{-1, 1\}^d$ be a $d/4$ -packing of the hypercube in ℓ_0 -norm. For a given $\delta \in (0, 1/4]$, define the risk function

$$R_v(\theta) := \frac{c}{d} \sum_{j=1}^d \frac{1 + \delta v_j}{2} |\theta - r| + \frac{1 - \delta v_j}{2} |\theta + r| = \frac{c}{d} \left(\frac{1 + \delta}{2} \|\theta - rv\|_1 + \frac{1 - \delta}{2} \|\theta + rv\|_1 \right). \quad (27)$$

By construction, whenever Θ contains the ℓ_∞ ball of radius r , this risk function has the unique minimizer

$$\theta_v^* := \operatorname{argmin}_{\theta \in \Theta} R_v(\theta) = rv \in r\{-1, 1\}^d \subset \Theta.$$

The risk (27) can be realized as the expectation of the median loss function

$$\ell(X, \theta) = \frac{c}{d} \|X - \theta\|_1, \quad (28a)$$

and a sampling scheme of the form

$$\text{Choose } X \in r\{-1, 1\}^d \text{ with independent coordinates, where } X_j = \begin{cases} r & \text{w.p. } \frac{1+\delta v_j}{2} \\ -r & \text{w.p. } \frac{1-\delta v_j}{2}. \end{cases} \quad (28b)$$

With these choices, it is straightforward to verify that $R_v(\theta) = \mathbb{E}[\ell(X, \theta)]$. The following lemma, due to Agarwal et al. [1], captures the separation properties of the collection $\{R_v\}_{v \in \mathcal{V}}$ of risk functionals:

Lemma 4. *Assume that Θ contains $[-r, r]^d$ and let R_v be defined by the risk (27). If $v, w \in \mathcal{V}$ with $v \neq w$, the discrepancy $\rho(R_v, R_w) \geq rc\delta/2$.*

As a final remark, for random variables $X \in \mathbb{R}^d$, the loss function (28a) is Lipschitz continuous (for appropriate choice of c) for *any* distribution P on X . Specifically, defining the $\operatorname{sign}(\cdot)$ function coordinate-wise, we have the subgradient equality $\partial\ell(x, \theta) = (c/d) \operatorname{sign}(\theta - x)$. Thus, choosing p, q to satisfy $1/q + 1/p = 1$ and $c = Ld^{1/q}$ yields a member of the collection of (L, p) -loss functions.

5.3 Mutual information bounds and hypothesis testing

We now return the main thread of the proof. From inspection of Fano's inequality (20), we see that it involves two quantities: (i) the log cardinality $\log |\mathcal{V}|$, and (ii) suitable upper bounds on the mutual information term. Since the log cardinality was specified during construction of our loss families in the preceding section, it remains to address the latter sub-problem.

Recall that Z_1, \dots, Z_n are unbiased subgradient estimates of the loss $\theta \mapsto \ell(X_i, \theta)$, where X_i are independent samples according to a distribution $P(\cdot | V)$. We assume that the samples Z_i are conditionally independent of V given X_i and the parameters θ ; this assumption is natural since Z is a random function of $\partial \ell(X_i, \theta)$. Our goal is to upper bound the mutual information between the sequence Z_1, \dots, Z_n of observed (stochastic) gradients and the random element $V \in \mathcal{V}$.

From Propositions 1 and 2, we know that the channel distributions Q guaranteeing privacy are supported on a finite set: in the case of $p = 1$, on (a multiple) of the standard basis vectors $\{\pm e_i\}_{i=1}^d$, and for $p = \infty$, on (a multiple of) the corners of the hypercube $\{-1, 1\}^d$. Thus (using the chain rule for mutual information [8]) we have the decomposition

$$I(Z_1, \dots, Z_n; V) = \sum_{i=1}^n [H(Z_i | Z_1, \dots, Z_{i-1}) - H(Z_i | V, Z_1, \dots, Z_{i-1})].$$

Let θ_i denote the point at which the i th gradient is computed. Then by inspection, we must have $\theta_i \in \sigma(Z_1, \dots, Z_{i-1})$. Since Z_i is conditionally independent of Z_1, \dots, Z_{i-1} given V and θ_i and conditioning decreases entropy, we have

$$\begin{aligned} H(Z_i | Z_1, \dots, Z_{i-1}) - H(Z_i | V, Z_1, \dots, Z_{i-1}) &= H(Z_i | Z_1, \dots, Z_{i-1}) - H(Z_i | V, \theta_i) \\ &\leq H(Z_i | \theta_i) - H(Z_i | V, \theta_i) \\ &= I(Z_i; V | \theta_i). \end{aligned}$$

In particular, letting F_i denote the distribution of θ_i , we have

$$I(Z_1, \dots, Z_n; V) \leq \sum_{i=1}^n \int_{\Theta} I(Z_i; V | \theta) dF_i(\theta) \leq \sum_{i=1}^n \sup_{\theta \in \Theta} I(Z_i; V | \theta). \quad (29)$$

We now state four lemmas, each bounding the mutual information between observed subgradients Z_i and the random variable V , for different choices of loss function ℓ and conditional distribution Q . The proof of each lemma begins by using the bound (29) to reduce the problem to estimating the mutual information $I(Z; V | \theta)$ for a single randomized gradient sample Z . Then, careful calculation of the distribution of $Z | V$ yields the final inequalities. As the proofs are somewhat long and technical, we defer them to Appendix B.

Lemma 5. *Let V be drawn uniformly at random from $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Let X have the distribution (23) conditional on $V = v$ and assume $\ell(X, \theta) = \langle X, \theta \rangle$. Let Z be constructed according to the conditional distribution specified by Proposition 1 given a subgradient $\partial \ell(X_i; \theta)$ with $Z \in [-M_\infty, M_\infty]^d$, where $M_\infty \geq c/d$. Then*

$$I(Z_1, \dots, Z_n; V) \leq n \frac{\delta^2 c^2}{M_\infty^2 d^2}.$$

See Appendix B.1 for a proof of Lemma 5.

Lemma 6. Let V be drawn uniformly at random from a set $\mathcal{V} \subset \{-1, 1\}^d$. Define the distribution $P(\cdot | V)$ on X to be such that the j th coordinate $X_j = rV_j$ with probability $(1+\delta)/2$ and $X_j = -rV_j$ with probability $(1-\delta)/2$, each coordinate independent of the others, where $r > 0$ is a constant. Let the loss function $\ell(X, \theta)$ be given as follows:

$$\ell(X, \theta) = \frac{c}{d} \|\theta - X\|_1.$$

Let Z be constructed according to the distribution specified by Proposition 1 conditional on a sub-gradient $\partial \ell(X_i; \theta)$, where $Z \in [-M_\infty, M_\infty]^d$ and $M_\infty \geq c/d$. Then

$$I(Z_1, \dots, Z_n; V) \leq n \frac{\delta^2 c^2}{M_\infty^2 d}.$$

See Appendix B.2 for a proof of Lemma 6.

Lemma 7. Let V be drawn uniformly at random from a set $\mathcal{V} \subset \{-1, 1\}^d$. Define the distribution $P(\cdot | A)$ on X as in the random sampling scheme (25) and use the loss (24). Let Z be constructed according to the conditional distribution specified by Proposition 2, where $Z \in \{z \in \mathbb{R}^d : \|z\|_1 \leq M_1\}$. Define $M = M_1/c$ and the constants

$$\gamma := \log \left(\frac{2d - 2 + \sqrt{(2d - 2)^2 + 4(M^2 - 1)}}{2(M - 1)} \right) \quad \text{and} \quad \Delta(\gamma) := \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d - 1)}.$$

Then

$$I(Z_1, \dots, Z_n; V) \leq n \delta^2 \Delta(\gamma)^2$$

We provide the proof of the lemma in Appendix B.3.

Lemma 8. Let V be drawn uniformly at random from $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Let $X | V$ be sampled according to the distribution (23), and let $Z | X = x$ have support on $\{-1, 1\}^d$ and have p.m.f.

$$q(z | x) \propto \begin{cases} \exp(\alpha) & \text{if } z^\top x > k \\ 1 & \text{if } z^\top x \leq k \end{cases}$$

for some $k \geq 0$. Define the constants $C_d(k)$ and $\Delta(\delta, \alpha, d, k)$ by

$$C_d(k) := \text{card} \left\{ z \in \{-1, 1\}^d : \langle z, x \rangle > k \right\} = \sum_{i=0}^{\lceil (d-k)/2 \rceil - 1} \binom{d}{i}.$$

and

$$\Delta(\delta, \alpha, d, k) := \delta \cdot \frac{e^\alpha - 1}{(e^\alpha + 1)C_d(k) + 2^d} \left(\binom{d-1}{\lceil (d-k)/2 \rceil - 1} \right).$$

Then

$$I(Z; V) \leq \Delta(\delta, \alpha, d, k)^2.$$

We provide the proof of the lemma in Appendix B.4.

In the proof of Theorems 1 and 2, we require one additional result for the cases when the dimension d is small; we apply the result instead of Fano's inequality. Specifically, we use Le Cam's method [30, 46], which provides lower bounds on the probability of error in binary hypothesis testing problems. In this setting, assume that $\mathcal{V} = \{-1, 1\}$ has two elements, and let $V \in \mathcal{V}$ be chosen uniformly at random from \mathcal{V} . If a procedure observes random variables Z_1, \dots, Z_n distributed according to Q_1^n if $V = 1$ and Q_{-1}^n if $V = -1$, then any estimate \hat{v} of V satisfies the lower bound

$$\mathbb{P}(\hat{v}(Z_1, \dots, Z_n) \neq V) \geq \frac{1}{2} - \frac{1}{2} \|Q_1^n - Q_{-1}^n\|_{\text{TV}}. \quad (30)$$

See, for example, Yu [46, Lemma 1] and Le Cam [30, Section 2]. Moreover, we have the following lemma.

Lemma 9. *Let Q_1 and Q_{-1} be distributions on $\{-1, 1\}$, where*

$$Q_1(Z = z) = \frac{1}{2} + \frac{1}{2} \cdot \begin{cases} \delta & \text{if } z = 1 \\ -\delta & \text{otherwise} \end{cases} \quad \text{and} \quad Q_{-1}(Z = z) = \frac{1}{2} + \frac{1}{2} \cdot \begin{cases} -\delta & \text{if } z = 1 \\ \delta & \text{otherwise} \end{cases}.$$

Let Q_i^n denote the n -fold product distribution of Q_i . Then for $\delta \in [0, 1/3]$,

$$\|Q_1^n - Q_{-1}^n\|_{\text{TV}} \leq \delta \sqrt{(3/2)n}.$$

We provide the proof of the lemma in Appendix B.5.

Equipped with these auxiliary results, we are now ready to prove our main theorems.

5.4 Proof of Theorem 1

We break the proof of Theorem 1 into three parts. In the first, we prove part (a) of the theorem assuming that the dimension $d \geq 9$. Next, we show part (a) for smaller values of the dimension, which requires Le Cam's bounding technique (30). Finally, we prove part (b). Roughly, our strategy is to apply Lemma 1 and one of Lemmas 5 or 6 to achieve a lower bound on the rate of convergence of any estimation procedure. We first recall the beginning of the previous section, stating the following application of Lemma 1 and Fano's inequality (20):

$$\frac{2}{\rho^*(\mathcal{V})} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \mathbb{P}_{P,Q}(\hat{v}(\mathcal{M}) \neq V) \geq 1 - \frac{I(Z_1, \dots, Z_n; V) + \log 2}{\log |\mathcal{V}|}. \quad (31)$$

Now we give the proof of the first statement of the theorem in the case that $d \geq 9$. Applying Lemmas 4 and 6, we immediately have the following specialization of the inequality (31):

$$\frac{4}{rc\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log |\mathcal{V}|} - n \frac{\delta^2 c^2}{M_\infty^2 d \log |\mathcal{V}|}.$$

Taking the set $\mathcal{V} \subset \{-1, 1\}^d$ to be a $d/4$ packing of the hypercube $\{-1, 1\}^d$ satisfying $|\mathcal{V}| \geq \exp(d/8)$, as described in Sections 5.2.2 and 5.2.3, we see that

$$\frac{4}{rc\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{8 \log 2}{d} - n \frac{8 \delta^2 c^2}{M_\infty^2 d^2}.$$

By the remarks following Lemma 4, we may take $L = c/d$. The numerical inequality $8 \log 2 < 6$ coupled with the preceding bound implies

$$\frac{4}{rdL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] > 1 - \frac{6}{d} - 8n \frac{\delta^2 L^2}{M_\infty^2}.$$

By our assumption that $d \geq 9$, if we choose $\delta = M_\infty/8L\sqrt{n}$, then we are guaranteed the lower bound $\frac{4}{rdL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] > \frac{1}{5}$, or equivalently

$$\mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] > \frac{rdL\delta}{20} = \frac{1}{160} \cdot \frac{M_\infty rd}{\sqrt{n}}.$$

When $d < 9$, we may reduce to the case that $d = 1$, since a lower bound in this setting extends to higher dimensions (though we may lose dimension dependence). For this case, we use the packing set $\mathcal{V} = \{-1, 1\}$ with the linear loss function from Lemma 2, which has $\rho^*(\mathcal{V}) = Lr\delta$. In this case, the marginal distribution $Q(\cdot | V)$ is given by

$$Q(Z = z | V = 1) = \frac{1}{2} + \begin{cases} \frac{\delta L}{2M} & \text{if } z = M \\ -\frac{\delta L}{2M} & \text{otherwise, i.e. if } z = -M. \end{cases}$$

Now, let $Q^n(\cdot | V)$ denote the distribution of Z_1, \dots, Z_n conditional on V . Then applying Lemma 1 and Le Cam's lower bound (30), we obtain the inequality

$$\frac{2}{rL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \mathbb{P}_{P,Q} (\hat{v}(\mathcal{M}) \neq V) \geq \frac{1}{2} - \frac{1}{2} \|Q^n(\cdot | V = 1) - Q^n(\cdot | V = -1)\|_{\text{TV}}.$$

By inspection, the distributions Q^n place us precisely in the conditions Lemma 9 specifies, so if $\delta \leq M/(3L)$, we have the bound

$$\frac{2}{rL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - \frac{\sqrt{3n}}{2\sqrt{2}} \cdot \frac{\delta L}{M}. \quad (32)$$

Multiplying both sides by $rL\delta$, then setting $\delta = M/(3L\sqrt{n}) \leq M/(3L)$, we have

$$\mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{(3\sqrt{2} - \sqrt{3})rM}{36\sqrt{2n}} \geq \frac{rM}{20.3\sqrt{n}}.$$

In turn, for any $d \leq 8$, we immediately find that $1/20.3 \geq d/163$, which completes the proof of Theorem 1(a).

For the second statement of the theorem, we use the linear losses of Section 5.2.1 and apply Lemmas 2 and 5 with the choice $\mathcal{V} = \{\pm e_i\}_{i=1}^d$. Since we are in the (L, ∞) -Lipschitz class of loss functions, we take $c = Ld$ in the sampling scheme (23). In this case, the lower bound (31) and Lemma 2's separation guarantee imply that

$$\frac{2}{Lr\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log(2d)} - \frac{I(Z_1, \dots, Z_n; V)}{\log(2d)}.$$

By assumption that $d \geq 2$, we have $\log 2 / \log(2d) \leq 1/2$, which, after an application of Lemma 5, yields

$$\frac{2}{Lr\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - n \frac{\delta^2 L^2}{M_\infty^2 \log(2d)}.$$

If we choose $\delta = M_\infty \sqrt{\log(2d)}/2L\sqrt{n}$, we see that we have

$$\frac{2}{Lr\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{4},$$

which is equivalent in this case to

$$\mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{rL\delta}{8} = \frac{1}{16} \cdot \frac{M_\infty r \sqrt{\log(2d)}}{\sqrt{n}}.$$

5.5 Proof of Theorem 2

The proof of Theorem 2 is quite similar to that of Theorem 1, except that we apply Lemma 7 in place of Lemmas 5 or 6. Indeed, following identical steps to those in the proof of Theorem 1, we see that with the packing $\mathcal{V}\{-1, 1\}^d$ of size $|\mathcal{V}| \geq \exp(d/8)$, we have

$$\begin{aligned} \frac{4}{rL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] &\geq 1 - \frac{\log 2}{\log |\mathcal{V}|} - n \frac{\delta^2 \Delta(\gamma)^2}{\log |\mathcal{V}|} \\ &\geq 1 - \frac{6}{d} - 8n \frac{\delta^2 \Delta(\gamma)^2}{d}. \end{aligned}$$

Consequently, if we choose $\delta = \sqrt{d}/(8\Delta(\gamma)\sqrt{n})$, then for all $d \geq 9$, we have the lower bound $\frac{4}{rL\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{5}$, or equivalently

$$\mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{rL\delta}{20} = \frac{1}{160} \cdot \frac{rL\sqrt{d}}{\sqrt{n}\Delta(\gamma)},$$

which completes the proof (as the case $d \leq 8$ is identical to that in Theorem 1).

5.6 Proof of Theorem 3

Since our optimization domain $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, we proceed similarly to our proof of Theorem 1 and use the linear losses of Section 5.2.1. Indeed, using the packing set $\mathcal{V} = \{\pm e_i\}_{i=1}^d$, we find that

$$\frac{2}{Lr\delta} \mathbb{E}_{P,Q} [\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq 1 - \frac{\log 2}{\log(2d)} - \frac{I(Z_1, \dots, Z_n; V)}{\log(2d)}$$

as earlier. For any $\alpha \leq 5/4$, we have $e^\alpha - 1 \leq 2\alpha$, and by properties of binomial coefficients and Stirling's approximation we have

$$\frac{1}{2^d} \binom{d-1}{\lceil (d-k)/2 \rceil - 1} \leq \frac{1}{2^d} \binom{d-1}{\lceil d/2 \rceil - 1} \leq \frac{1}{\sqrt{d}}$$

for any k . Now, for any distribution Q satisfying optimal local differential privacy at a differential privacy level α , Proposition 3 implies Q is a convex combination of distributions with p.m.f.s of the form in Lemma 8. Applying the convexity of mutual information—taking a convex combination of channel distributions Q can only reduce mutual information—and Lemma 8, we thus obtain

$$\begin{aligned} I(Z_1, \dots, Z_n; V) &\leq n \max_{k \geq 0} \Delta(\delta, \alpha, d, k)^2 \\ &\leq n\delta^2(e^\alpha - 1)^2 \max_k \left(\frac{1}{(e^\alpha + 1)C_d(k) + 2^d} \binom{d-1}{\lceil (d-k)/2 \rceil - 1} \right)^2 \leq 4n\delta^2\alpha^2 \cdot \frac{1}{d}. \end{aligned}$$

As a consequence, we have the lower bound

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{2} - \max_k \frac{n\Delta(\delta, \alpha, d, k)^2}{\log(2d)} \geq \frac{1}{2} - \frac{4n\delta^2\alpha^2}{d\log(2d)}.$$

By choosing $\delta = \sqrt{d\log(2d)}/4\alpha\sqrt{n}$, we find that

$$\frac{2}{Lr\delta}\mathbb{E}_{P,Q}[\epsilon_n(\mathcal{M}, \ell, \Theta, P)] \geq \frac{1}{4},$$

which is equivalent to the bound given in the theorem.

5.7 Proof of Corollary 1

Since $\Theta \subseteq \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq r\}$, the bound (14a) guarantees that mirror descent obtains convergence rate $\mathcal{O}(M_\infty r \sqrt{\log d}/\sqrt{n})$. This matches the second statement of Theorem 1. Now fix our desired amount of mutual information I^* . From the remarks following Proposition 1, if we must guarantee that $I^* \geq \sup_P I(P, Q)$ for any distribution P and loss function ℓ whose gradients are bounded in ℓ_∞ -norm by L , we *must* (because of the uniqueness of the optimal privacy distribution Q) have

$$I^* \asymp \frac{dL^2}{M_\infty^2}. \quad (33)$$

Up to higher order terms, to guarantee a level of privacy with mutual information I^* , we must allow gradient noise up to a level $M_\infty = L\sqrt{d/I^*}$. The equality (33) establishes that for a given level of allowed mutual information I^* , if optimal local privacy holds, then we must have $M_\infty \asymp L\sqrt{d}/\sqrt{I^*}$. That is, we have a bijection between I^* and M_∞ whenever optimal local privacy holds, so substituting $M_\infty = L\sqrt{d}/\sqrt{I^*}$ into our upper and lower bounds gives the corollary. \square

5.8 Proof of Corollary 2

According to the conditions of optimal local privacy, if we must guarantee that $I^* \geq \sup_P I(P, Q)$ for any loss function ℓ whose gradients are bounded in ℓ_1 -norm by L , we must have

$$I^* \asymp \frac{dL^2}{2M_1^2},$$

using Corollary 4 after the statement of Proposition 2. Rewriting this, we see that we must have $M_1 = L\sqrt{d/2I^*}$ (to higher order terms) to be able to guarantee an amount of privacy I^* . As in the ℓ_∞ case, we have a bijection between the multiplier M_1 and the amount of information I^* and can apply similar techniques. Now recall the convergence guarantee (14b) provided by stochastic gradient descent. Since the ℓ_∞ -ball of radius r is contained in the ℓ_2 -ball of radius $r_2 = r\sqrt{d}$, and $\|g\|_1 \leq \|g\|_2$ for all $g \in \mathbb{R}^d$, stochastic gradient descent guarantees that $\epsilon_n^*(\mathfrak{L}, \Theta) \leq CM_1 r \sqrt{d}/\sqrt{n}$. Applying the lower bound provided by Theorem 2 and substituting for M_1 completes the proof. \square

5.9 Proof of Corollary 3

Without loss of generality (by scaling), we assume that $L = 1$. Now we consider Proposition 3, which characterizes the distributions satisfying optimal local differential privacy. We use the proposition

to find an upper bound on M_∞ in terms of the differential privacy level α , which in turn allows us to apply the bound from mirror descent (14a). Instead of directly using Proposition 3, it is simpler to use the linear program (52) in its proof, and note that finding a lower bound on t (in the LP) as a function of α provides an upper bound on M_∞ since $M_\infty = 1/t$. Now, in the linear program (52), we choose the values for $q(z)$ specified by Lemma 16. Let q_+ and q_- denote the larger and smaller probabilities, respectively. Fix an $x \in \{-1, 1\}^d$, and let z range over $\{-1, 1\}^d$. With those choices, we note that for d odd,

$$\begin{aligned} \sum_{z: \langle z, x \rangle > 0} z &= \sum_{z: \langle z, x \rangle = 1} z + \sum_{z: \langle z, x \rangle = 3} z + \dots + \sum_{z: \langle z, x \rangle = d} z \\ &= \left[\binom{d-1}{\frac{d-1}{2}} - \binom{d-1}{\frac{d+1}{2}} \right] x + \left[\binom{d-1}{\frac{d+1}{2}} - \binom{d-1}{\frac{d+3}{2}} \right] x + \dots = \binom{d-1}{\frac{d-1}{2}} x. \end{aligned}$$

For d even, a similar calculation yields $\sum_{z: \langle z, x \rangle > 0} z = \binom{d-1}{d/2} x$. As a consequence, we find that

$$\sum_z z q(z | x) = q_+ \sum_{z: \langle z, x \rangle > 0} z + q_- \sum_{z: \langle z, x \rangle \leq 0} z = x(q_+ - q_-) \cdot \begin{cases} \binom{d-1}{\frac{d-1}{2}} & d \text{ odd} \\ \binom{d-1}{d/2} & d \text{ even.} \end{cases}$$

Focusing on the odd case for simplicity—identical bounds hold in the even case—we have for a universal constant $c > 0$ that

$$(q_+ - q_-) \binom{d-1}{\frac{d-1}{2}} = \frac{e^\alpha - 1}{2^{d-1}(e^\alpha + 1)} \binom{d-1}{\frac{d-1}{2}} \geq c \frac{e^\alpha - 1}{e^\alpha + 1} \frac{1}{\sqrt{d}} \geq c \frac{\alpha}{\sqrt{d}},$$

the first inequality following from Stirling's approximation and the second from convexity of the function $\alpha \mapsto e^\alpha$. In particular, we see that the minimizing value t in the linear program (52) will satisfy $t \geq c\alpha/\sqrt{d}$, which in turn yields $M_\infty = 1/t \leq \sqrt{d}/(c\alpha)$. Noting that the lower bound in the corollary is given by Theorem 3, applying the convergence guarantee (14a) of mirror descent based on M_∞ completes the proof. \square

6 Discussion

We have studied methods for protecting privacy in general statistical risk minimization problems, in particular techniques that maintain privacy between the data X_1, \dots, X_n and the estimation method \mathcal{M} . As a consequence of our focus, we were able to provide a general technique for obtaining sharp tradeoffs between privacy protection and estimation rates, which are a natural measure of utility for statistical problems.

We believe that there are a number of remaining open issues and areas for future work. First, we studied procedures that access each datum only once, and through a perturbed view Z_i of the subgradient $\partial \ell(X_i, \theta)$, which allowed us to use (essentially) arbitrary convex losses. A natural question is whether there are restrictions of the class of loss functions so that a transformed version (Z_1, \dots, Z_n) of the data are sufficient for inference. For instance, Zhou et al. [47, 48] study applications in which a data matrix $X = [X_1 \ \dots \ X_n]^\top \in \mathbb{R}^{n \times d}$ is pre-multiplied by a normal matrix $\Phi \in \mathbb{R}^{m \times n}$, where $m \ll n$, and statistical inference is performed using ΦX . For problems such as linear regression and PCA, the resulting estimators enjoy good statistical properties. This transformation, however, cannot be computed without the entire dataset at one's disposal. Nonparametric

data releases, such as those studied by Hall et al. [22], could provide insights here, though again, current approaches require the data to be aggregated by a trusted curator before release.

Our constraints on the privacy-inducing channel distribution Q require that its support lie in some compact set. We find this restriction useful, but perhaps it possible to achieve faster estimation rates if all we require are moment conditions, for example, $\mathbb{E}_Q[\|Z - X\|_p^2 | X] \leq M^2$. A better understanding of general privacy-preserving channels Q for alternative constraints to those we have proposed is also desirable. Moreover, one might consider attempting only to guarantee that $\phi(X)$ is private, where ϕ is some (known) function. For example, members of a dataset may not care if their genders are known, but more personal features of X may be more sensitive.

These questions do not appear to have easy answers, especially when we wish to allow each provider of a single datum to be able to guarantee his or her own privacy. Nevertheless, we hope that our view of privacy and the techniques we have developed herein prove fruitful, and we hope to investigate some of the above issues in future work.

References

- [1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, May 2012.
- [2] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.
- [3] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., 1989.
- [4] P. Billingsley. *Probability and Measure*. Wiley, Second edition, 1986.
- [5] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the Fourtieth Annual ACM Symposium on the Theory of Computing*, 2008.
- [6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [7] K. Chaudhuri, C. Moneleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.
- [8] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.
- [9] L. H. Cox, A. F. Karr, and S. K. Kinney. Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. *International Statistical Review*, 79(2):160–199, 2011.
- [10] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.
- [11] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 202–210, 2003.
- [12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.
- [13] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *Journal of the American Statistical Association*, 81(393):10–18, 1986.
- [14] G. T. Duncan and D. Lambert. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7(2):207–217, 1989.

- [15] C. Dwork. Differential privacy: a survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 2008.
- [16] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on the Theory of Computing*, 2009.
- [17] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.
- [18] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second Symposium on Principles of Database Systems*, pages 211–222, 2003.
- [19] I. P. Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1972.
- [20] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the Forty-First Annual ACM Symposium on the Theory of Computing*, 2009. URL <http://theory.stanford.edu/~tim/papers/priv.pdf>.
- [21] R. M. Gray. *Entropy and Information Theory*. Springer, 1990.
- [22] R. Hall, A. Rinaldo, and L. Wasserman. Random differential privacy. URL <http://arxiv.org/abs/1112.2680>, 2011.
- [23] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the Forty-Second Annual ACM Symposium on the Theory of Computing*, pages 705–714, 2010.
- [24] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I & II*. Springer, 1996.
- [25] O. Kallenberg. *Foundations of Modern Probability*. Springer, 1997.
- [26] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232, 2006.
- [27] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.
- [28] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association for Computing Machinery*, 45(6):983–1006, 1998.
- [29] L. Le Cam. On the asymptotic theory of estimation and hypothesis testing. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 129–156, 1956.
- [30] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Annals of Statistics*, 1(1):38–53, 1973.
- [31] Y. Liang, H. V. Poor, and S. Shamai. Information theoretic security. *Foundations and Trends in Communications and Information Theory*, 5(4):355–580, 2008.
- [32] O. L. Mangasarian. Uniqueness of solution in linear programming. *Linear Algebra and its Applications*, 25:151–162, 1979.
- [33] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [34] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [35] R. R. Phelps. *Lectures on Choquet’s Theorem, Second Edition*. Springer, 2001.
- [36] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM*

- Journal on Control and Optimization*, 30(4):838–855, 1992.
- [37] J. P. Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100:1103–1113, 2005. URL <http://sisla06.samsi.info/ndhs/dc/Papers/JerryJasa05.pdf>.
 - [38] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *Journal of Privacy and Confidentiality*, To appear, 2011. URL <http://arxiv.org/abs/0911.5708v1>.
 - [39] L. Sankar, S. R. Rajagopalan, and H. V. Poor. An information-theoretic approach to privacy. In *The 48th Allerton Conference on Communication, Control, and Computing*, pages 1220–1227, 2010.
 - [40] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing*, 2011.
 - [41] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.
 - [42] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.
 - [43] S. L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.
 - [44] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.
 - [45] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.
 - [46] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.
 - [47] S. Zhou, J. Lafferty, and L. Wasserman. Compressed regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.
 - [48] S. Zhou, K. Ligett, and L. Wasserman. Differential privacy with compression. In *Proceedings of the 2009 IEEE International Symposium on Information Theory*, 2009.
 - [49] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A Unbiasedness

In this appendix, we show that if an optimization procedure receives biased subgradients it is possible to be arbitrarily wrong. In particular, we do so by constructing a simple problem instance. Fix a bias $b > 0$ and consider the following one dimensional problem:

$$\text{minimize } f(\theta) := \frac{b\theta}{2} \quad \text{subject to } \theta \in [-c, c].$$

If a gradient oracle returns biased gradients of the form $-b/2$ at each point $\theta \in [-c, c]$, it is impossible to distinguish the objective from $-b\theta/2$. The minimizer of this objective is $\theta_{\text{bias}} = \text{sign}(b)c$. The true optimal point is $\theta^* = -\text{sign}(b)c$, yielding the worst possible error

$$f(\theta_{\text{bias}}) - f(\theta^*) = \sup_{\theta \in [-c, c]} f(\theta) - \inf_{\theta \in [-c, c]} f(\theta).$$

We can show this more formally using an information theoretic derivation similar to that in Section 5. Omitting details, the argument is as follows. In the notation of Section 5, if a bias is chosen independently of the parameters $v \in \mathcal{V}$ of the risk R_v , then there is a *bounded* amount of mutual information that can be communicated to any optimization procedure. Consequently, Fano's inequality (20) guarantees that the estimation accuracy of any procedure must be bounded away from zero.

B Calculation of the Mutual Information for Sampling Strategies

This appendix is devoted to the proofs of Lemma 5, Lemma 6, and Lemma 7. The proofs of the latter two require a minor lemma, which we present here before giving the proofs proper.

Lemma 10. *Let $1 > p > \delta > 0$ and $p + \delta \leq 1$. Then*

$$(p + \delta) \log(p + \delta) + (p - \delta) \log(p - \delta) > 2p \log p.$$

Proof Since the function $p \mapsto f(p) = p \log p$ is strictly convex over $[0, \infty)$, we may apply convexity. Indeed, $p = \frac{1}{2}(p + \delta) + \frac{1}{2}(p - \delta)$, so

$$p \log p = f\left(\frac{1}{2}(p + \delta) + \frac{1}{2}(p - \delta)\right) < \frac{1}{2}f(p + \delta) + \frac{1}{2}f(p - \delta),$$

which is the desired result. □

B.1 Proof of Lemma 5

It is clear that the subgradient set $\partial \ell(X_i; \theta)$ is independent of θ , so we may use the inequality (29) to bound the mutual information of V and a single sample Z . Define $M = M_\infty d/c$. Since the sampling scheme (23) is independent per-coordinate, we see immediately that if Z_j denotes the j th coordinate of Z then

$$I(Z; V) = H(Z) - H(Z | V) \leq d \log(2) - \sum_{j=1}^d H(Z_j | V).$$

Since V is uniformly chosen from one of $2d$ vectors, we additionally find that

$$I(Z; V) \leq d \left[\log 2 - \frac{1}{2d} \sum_{v \in \mathcal{V}} H(Z | V = v) \right].$$

By the choice of our sampling scheme for X and Z , we see that $H(Z | V = v)$ is identical for each $v \in \mathcal{V}$, and we have

$$Q(Z_j = M_\infty | V_j = v_j = 0) = \frac{1}{2}, \quad \text{and} \quad Q(Z_j = -M_\infty | V_j = v_j = 0) = \frac{1}{2}.$$

On the other hand, by our choice of sampling scheme, for the “on” index in V , we have

$$\begin{aligned} Q(Z_j = M_\infty \mid V_j = v_j = 1) &= Q(Z_j = M_\infty \mid X_j = c/d)P(X_j = c/d \mid V_j = v_j = 1) \\ &\quad + Q(Z_j = M_\infty \mid X_j = -c/d)P(X_j = -c/d \mid V_j = v_j = 1) \\ &= \left(\frac{M+1}{2M}\right) \left(\frac{1+\delta}{2}\right) + \left(\frac{M-1}{2M}\right) \left(\frac{1-\delta}{2}\right) = \frac{1}{2} + \frac{\delta}{2M}. \end{aligned}$$

Consequently, defining the Bernoulli entropy $h(p) = -p \log p - (1-p) \log(1-p)$, then

$$\begin{aligned} I(Z; V) &\leq d \left[\log 2 - \frac{1}{2d} \left((2d-2) \log 2 + 2h\left(\frac{1}{2} + \frac{\delta}{2M}\right) \right) \right] \\ &= \log 2 + \left(\frac{1}{2} + \frac{\delta}{2M}\right) \log \left(\frac{1}{2} + \frac{\delta}{2M}\right) + \left(\frac{1}{2} - \frac{\delta}{2M}\right) \log \left(\frac{1}{2} - \frac{\delta}{2M}\right). \end{aligned}$$

The concavity of the function $p \mapsto \log(p)$ yields that $\log(1/2 + p) \leq \log(1/2) + 2p$, so

$$I(Z; V) \leq \log 2 + \left(\frac{1}{2} + \frac{\delta}{2M}\right) \left(-\log 2 + \frac{\delta}{M}\right) + \left(\frac{1}{2} - \frac{\delta}{2M}\right) \left(-\log 2 - \frac{\delta}{M}\right) = \frac{\delta^2}{M^2}.$$

Making the substitution $M = M_\infty d/c$ completes the proof.

B.2 Proof of Lemma 6

By using the inequality (29), a bound on the mutual information $I(Z; V \mid \theta)$ implies a bound on the joint information in the statement of the lemma, so we focus on bounding the mutual information of a single sample Z . In addition, it is no loss of generality to assume that $r = 1$.

Define $M = M_\infty d/c$ to be the multiple of the ℓ_∞ -norm of the subgradients that we take, and let Z_j denote the j th coordinate of Z . Using the coordinate-wise independence of the sampling, we have

$$I(Z; V \mid \theta) = H(Z \mid \theta) - H(Z \mid V, \theta) \leq d \log(2) - \sum_{j=1}^d H(Z_j \mid V_j, \theta_j).$$

Now consider the distribution of Z_j given V_j and θ_j . By symmetry, the distribution has identical entropy for any value of V_j , so we may fix $V = v$ and assume $v_j = 1$ without loss of generality. Then for $\theta_j \in (-1, 1)$, the j th component of the subgradient $\partial \ell(X; \theta)$ is $-X_j$, whence we see that

$$\begin{aligned} Q(Z_j = M_\infty \mid v_j = 1, \theta_j) &= Q(Z_j = M_\infty \mid X_j = 1, \theta_j)P(X_j = M_\infty \mid v_j = 1) + Q(Z_j = M_\infty \mid X_j = -1, \theta_j)P(X_j = -1 \mid v_j = 1) \\ &= \left(\frac{M-1}{2M}\right) \left(\frac{1+\delta}{2}\right) + \left(\frac{M+1}{2M}\right) \left(\frac{1-\delta}{2}\right) \\ &= \frac{2M-2\delta}{4M} = \frac{1}{2} - \frac{\delta}{2M}. \end{aligned}$$

Similarly, $Q(Z_j = -M_\infty \mid v_j = 1, \theta_j) = \frac{1}{2} + \frac{\delta}{2M}$. If $\theta_j \geq 1$, then we have that the subgradient $\partial|\theta_j - X_j| = 1$ with probability 1, and thus

$$Q(Z_j = M_\infty \mid v_j = 1, \theta_j) = \left(\frac{M+1}{2M}\right) \left(\frac{1+\delta}{2}\right) + \left(\frac{M+1}{2M}\right) \left(\frac{1-\delta}{2}\right) = \frac{1}{2},$$

which increases the entropy $H(Z_j | V_j, \theta_j)$ by Lemma 10. Thus we see that $\theta_j \in (-1, 1)$, yielding the Bernoulli marginal $(\frac{1}{2} + \delta/2M, \frac{1}{2} - \delta/2M)$ on $Z_j | V_j$, has the smallest entropy $H(Z_j | V_j, \theta_j)$. Summarizing, we have

$$I(Z; V | \theta) \leq d \log(2) + d \left[\left(\frac{1}{2} + \frac{\delta}{2M} \right) \log \left(\frac{1}{2} + \frac{\delta}{2M} \right) + \left(\frac{1}{2} - \frac{\delta}{2M} \right) \log \left(\frac{1}{2} - \frac{\delta}{2M} \right) \right].$$

As in the proof of Lemma 5, we use the concavity of log to see that

$$\begin{aligned} I(Z; V | \theta) &\leq d \log(2) + d \left[\left(\frac{1}{2} + \frac{\delta}{2M} \right) (-\log(2) + \delta/M) + \left(\frac{1}{2} - \frac{\delta}{2M} \right) (-\log(2) - \delta/M) \right] \\ &= d \left(\frac{1}{2} + \frac{\delta}{2M} \right) \left(\frac{\delta}{M} \right) + d \left(\frac{1}{2} - \frac{\delta}{2M} \right) \left(-\frac{\delta}{M} \right) = \frac{d\delta^2}{M^2}. \end{aligned}$$

Applying the bound (29) and replacing $M = M_\infty d/c$ completes the proof.

B.3 Proof of Lemma 7

Letting Z denote a single subgradient sample using the conditional distribution Q specified by Proposition 2, we first prove that

$$I(Z; V | \theta) \leq \delta^2 \Delta(\gamma)^2 \quad \text{for any } \theta \in \mathbb{R}^d. \quad (34)$$

Recall the SVM risk (26) defined using the individual hinge losses (24): by construction, whenever $X = e_i$, then the loss is equal to $c[r - \theta_i]_+$. We have

$$\partial \ell(e_i, \theta) = c \begin{cases} 0 & \text{if } \theta_i > r \\ -e_i & \text{otherwise} \end{cases} \quad \text{and} \quad \partial \ell(-e_i, \theta) = c \begin{cases} 0 & \text{if } \theta_i < -r \\ e_i & \text{otherwise.} \end{cases}$$

For the remainder of this proof, we use the shorthand

$$D_\gamma := e^\gamma + e^{-\gamma} + 2(d-2)$$

for the denominator in many of our expressions. By the construction in Proposition 2, we have

$$Q(Z = M_1 e_i | X = e_i, \theta) = \begin{cases} \frac{e^{-\gamma}}{D_\gamma} & \text{if } \theta_i \leq r \\ \frac{1}{2d} & \text{if } \theta_i > r, \end{cases} \quad (35)$$

and similarly we have for $j \neq i$ that

$$Q(Z = M_1 e_j | X = e_i, \theta) = \begin{cases} \frac{1}{D_\gamma} & \text{if } \theta_i \leq r \\ \frac{1}{2d} & \text{if } \theta_i > r. \end{cases} \quad (36)$$

For $X = -e_i$, we have the conditional distribution parallel to (35):

$$Q(Z = M_1 e_i | X = -e_i, \theta) = \begin{cases} \frac{e^\gamma}{D_\gamma} & \text{if } \theta_i \geq -r \\ \frac{1}{D_\gamma} & \text{if } \theta_i < -r. \end{cases}$$

For any given θ , we have that

$$I(Z; V | \theta) = H(Z | \theta) - H(Z | V, \theta) \leq \log(2d) - \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} H(Z | \theta, V = v) \quad (37)$$

since the choice of V is uniform and Z takes on at most $2d$ values. We thus use the conditional distributions (35) and (36) to compute the entropy $H(Z | \theta, V)$ (specifically, the minimal such entropy across all values of θ). To do this, we compute the marginal distribution $Q(z | v)$, arguing that $H(Z | \theta, V)$ is minimal for $\theta \in \text{int}[-r, r]^d$. When $\theta_j \in (-r, r)$ for all j , we have

$$\begin{aligned} Q(Z = M_1 e_i | V = v, \theta) &= \sum_{j=1}^d Q(Z = M_1 e_i | X = e_j, \theta) P(X = e_j | V = v) \\ &\quad + \sum_{j=1}^d Q(Z = M_1 e_i | X = -e_j, \theta) P(X = -e_j | V = v). \end{aligned}$$

When $v_i = 1$, we thus have that

$$\begin{aligned} Q(Z = M_1 e_i | V = v, \theta) &= \frac{1 + \delta e^{-\gamma}}{2d} \frac{1}{D_\gamma} + \frac{1 - \delta e^\gamma}{2d} \frac{1}{D_\gamma} + \sum_{j \neq i} \frac{1}{D_\gamma} \left(\frac{1 + \delta v_j}{2d} + \frac{1 - \delta v_j}{2d} \right) \\ &= \frac{e^\gamma + e^{-\gamma} + \delta(e^{-\gamma} - e^\gamma)}{2d D_\gamma} + \frac{d-1}{d D_\gamma} = \frac{1}{2d} + \frac{\delta(e^{-\gamma} - e^\gamma)}{2d D_\gamma}, \end{aligned} \quad (38a)$$

and under the same condition,

$$Q(Z = -M_1 e_i | A = v, \theta) = \frac{e^\gamma + e^{-\gamma} + \delta(e^\gamma + e^{-\gamma})}{2d D_\gamma} + \frac{d-1}{d D_\gamma} = \frac{1}{2d} + \frac{\delta(e^\gamma - e^{-\gamma})}{2d D_\gamma}. \quad (38b)$$

If for any (possibly multiple) indices j we have $\theta_j \notin (-r, r)$, then via a bit of algebra and the conditional distributions (35) and (36), we see that there exists an $\epsilon \in (0, 1)$ such that

$$Q(Z = M_1 e_i | V = v, \theta) = \epsilon \frac{1}{2d} + (1 - \epsilon) \left(\frac{1}{2d} + \frac{\delta(e^{-\gamma} - e^\gamma)}{2d D_\gamma} \right).$$

Lemma 10 then implies that if $\theta \in \text{int}[-r, r]^d$ while $\theta' \notin \text{int}[-r, r]^d$, then

$$H(Z | \theta, V = v) < H(Z | \theta', V = v).$$

Since we seek an upper bound on the mutual information, we may thus assume without loss of generality that $\theta \in \text{int}[-r, r]^d$.

Now we compute the entropy $H(Z | \theta, v)$ using the marginal conditional distributions (38a) and (38b), which describe $Z | V$ when $\theta \in \text{int}[-r, r]^d$. Indeed, recall the definition in the statement of the lemma of the difference $\Delta(\gamma)$. For $z \in \{\pm M_1 e_j\}_{j=1}^d$, define the relation $z \sim v$ to mean that if $z = M_1 e_i$, then $v_i = 1$ and if $z = -M_1 e_i$ then $v_i = -1$. We then see that the entropy is

$$\begin{aligned} H(Z | \theta, V = v) &= - \sum_{z \sim v} Q(z | v, \theta) \log Q(z | v, \theta) - \sum_{z \not\sim v} Q(z | v, \theta) \log Q(z | v, \theta) \\ &= -d \left(\frac{1}{2d} + \frac{\delta \Delta(\gamma)}{2d} \right) \log \left(\frac{1}{2d} + \frac{\delta \Delta(\gamma)}{2d} \right) - d \left(\frac{1}{2d} - \frac{\delta \Delta(\gamma)}{2d} \right) \log \left(\frac{1}{2d} - \frac{\delta \Delta(\gamma)}{2d} \right). \end{aligned}$$

As in the proofs of Lemmas 5 and 6, we use the concavity of $\log(\cdot)$ to see that

$$\begin{aligned} -H(Z \mid \theta, V = v) &= \left(\frac{1}{2} + \frac{\delta\Delta(\gamma)}{2}\right) \log\left(\frac{1}{2d} + \frac{\delta\Delta(\gamma)}{2d}\right) + \left(\frac{1}{2} - \frac{\delta\Delta(\gamma)}{2}\right) \log\left(\frac{1}{2d} - \frac{\delta\Delta(\gamma)}{2d}\right) \\ &\leq \left(\frac{1}{2} + \frac{\delta\Delta(\gamma)}{2}\right) (-\log(2d) + \delta\Delta(\gamma)) + \left(\frac{1}{2} - \frac{\delta\Delta(\gamma)}{2}\right) (-\log(2d) - \delta\Delta(\gamma)) \\ &= -\log(2d) + \delta^2\Delta(\gamma)^2. \end{aligned}$$

Invoking the earlier bound (37) and adding $\log(2d)$ to the above expression completes the proof of the claim (34).

B.4 Proof of Lemma 8

Let Z_j denote the j th coordinate of Z . We first argue that conditional on V , the random variable Z has independent coordinates. Indeed, let $q_+ = q(z \mid x)$ for z such that $z^\top x > k$ and $q_- = e^{-\alpha}q_+$. Without loss of generality, we may take $V = e_1$, the first basis vector, and hence

$$\begin{aligned} Q(Z = z \mid V = e_1) &= \sum_{x \in \{-1, 1\}^d} Q(Z = z \mid X = x) P(X = x \mid V = e_1) \\ &= \frac{1}{2^{d-1}} \sum_{x \in \{-1, 1\}^d} Q(Z = z \mid X = x) \cdot \frac{1 + x_1\delta}{2} \\ &= \frac{1}{2^{d-1}} \left[\sum_{x: \langle z, x \rangle > k} q_+ \frac{1 + x_1\delta}{2} + \sum_{x: \langle z, x \rangle \leq k} q_- \frac{1 + x_1\delta}{2} \right]. \end{aligned} \quad (39)$$

Now, if $z_1 = 1$, then

$$\sum_{x: \langle x, z \rangle > k} \frac{1 + x_1\delta}{2} = \sum_{x: \langle x, z \rangle > k, x_1 = 1} \frac{1 + \delta}{2} + \sum_{x: \langle x, z \rangle > k, x_1 = -1} \frac{1 + \delta}{2} = \frac{1 + \delta}{2} C_{d-1}(k-1) + \frac{1 - \delta}{2} C_{d-1}(k+1)$$

and similarly

$$\sum_{x: \langle x, z \rangle \leq k} \frac{1 + x_1\delta}{2} = \frac{1 + \delta}{2} (2^{d-1} - C_{d-1}(k-1)) + \frac{1 - \delta}{2} (2^{d-1} - C_{d-1}(k+1)).$$

On the other hand, we find that if $z_1 = -1$, then similar equalities hold, but with the counters $C_{d-1}(k-1)$ and $C_{d-1}(k+1)$ flipped:

$$\begin{aligned} \sum_{x: \langle x, z \rangle > k} \frac{1 + x_1\delta}{2} &= \frac{1 + \delta}{2} C_{d-1}(k+1) + \frac{1 - \delta}{2} C_{d-1}(k-1) \\ \sum_{x: \langle x, z \rangle \leq k} \frac{1 + x_1\delta}{2} &= \frac{1 + \delta}{2} (2^{d-1} - C_{d-1}(k+1)) + \frac{1 - \delta}{2} (2^{d-1} - C_{d-1}(k-1)). \end{aligned}$$

In particular, we find that so long as the first coordinate $z_1 = z'_1$ of z remains constant, then $Q(Z = z \mid V = e_1) = Q(Z = z' \mid V = e_1)$, and that we thus have Z_2, \dots, Z_d are distributed uniformly at random in $\{-1, 1\}^d$.

We now determine q_+ and compute the marginal value $Q(Z_1 = 1 \mid V = e_1)$. For the first, we note that

$$C_d(k)q_+ + (2^d - C_d(k))q_- = 1, \quad \text{or} \quad C_d(k)q_+ + e^{-\alpha}(2^d - C_d(k))q_+ = 1,$$

which yields the expressions

$$q_+ = \frac{e^\alpha}{(e^\alpha - 1)C_d(k) + 2^d} \quad \text{and} \quad q_- = \frac{1}{(e^\alpha - 1)C_d(k) + 2^d}.$$

By the expression (39) and calculations following, we thus find that when $z_1 = 1$, we have

$$\begin{aligned} q(z \mid e_1) &= \frac{1}{2^{d-1}} \cdot \left[q_+ \left(\frac{1+\delta}{2} C_{d-1}(k-1) + \frac{1-\delta}{2} C_{d-1}(k+1) \right) \right. \\ &\quad \left. + q_- \left(\frac{1+\delta}{2} (2^{d-1} - C_{d-1}(k-1)) + \frac{1-\delta}{2} (2^{d-1} - C_{d-1}(k+1)) \right) \right] \\ &= \frac{1}{2^{d-1}} \cdot \left[2^{d-1} q_- + \frac{1}{2} (q_+ - q_-) (C_{d-1}(k-1) + C_{d-1}(k+1)) \right. \\ &\quad \left. + \frac{\delta}{2} (q_+ - q_-) (C_{d-1}(k-1) - C_{d-1}(k+1)) \right], \end{aligned} \quad (40a)$$

and similarly when $z_1 = -1$ we have

$$\begin{aligned} q(z \mid e_1) &= \frac{1}{2^{d-1}} \cdot \left[2^{d-1} q_- + \frac{1}{2} (q_+ - q_-) (C_{d-1}(k-1) + C_{d-1}(k+1)) \right. \\ &\quad \left. - \frac{\delta}{2} (q_+ - q_-) (C_{d-1}(k-1) - C_{d-1}(k+1)) \right]. \end{aligned} \quad (40b)$$

Now note that

$$C_{d-1}(k-1) - C_{d-1}(k+1) = \sum_{i=0}^{\lceil (d-k)/2 \rceil - 1} \binom{d-1}{i} - \sum_{i=0}^{\lceil (d-k)/2 \rceil - 2} \binom{d-1}{i} = \binom{d-1}{\lceil (d-k)/2 \rceil - 1}$$

and that the difference

$$q_+ - q_- = \frac{e^\alpha - 1}{(e^\alpha - 1)C_d(k) + 2^d}.$$

Recalling the definition of the constant Δ , we thus find from the expansions (40a) and (40b)—since they must sum to 1—that

$$Q(Z = z \mid V = e_1) = \frac{1}{2^{d-1}} \cdot \begin{cases} \frac{1}{2} + \frac{\Delta(\delta, \alpha, d, k)}{2} & \text{if } z_1 = 1 \\ \frac{1}{2} - \frac{\Delta(\delta, \alpha, d, k)}{2} & \text{if } z_1 = -1. \end{cases} \quad (41)$$

It is clear that similar statements hold in the other symmetric cases (i.e. if $V = -e_2$, then the probabilities depend on $z_2 = -1$ or 1).

It remains to use the marginalized representation (41) to compute the bound on the mutual information in the statement of the lemma. To that end, note that

$$\begin{aligned}
I(Z; V) &= H(Z) - H(Z | V) \leq d \log 2 - \frac{1}{2d} \sum_v H(Z | V = v) \\
&= d \log 2 - (d-1) \log 2 \\
&\quad + \left(\frac{1}{2} + \frac{\Delta(\delta, \alpha, d, k)}{2} \right) \log \left(\frac{1}{2} + \frac{\Delta(\delta, \alpha, d, k)}{2} \right) + \left(\frac{1}{2} - \frac{\Delta(\delta, \alpha, d, k)}{2} \right) \log \left(\frac{1}{2} - \frac{\Delta(\delta, \alpha, d, k)}{2} \right) \\
&\leq \log 2 + \left(\frac{1}{2} + \frac{\Delta(\delta, \alpha, d, k)}{2} \right) \left[\log \frac{1}{2} + \Delta(\delta, \alpha, d, k) \right] + \left(\frac{1}{2} - \frac{\Delta(\delta, \alpha, d, k)}{2} \right) \left[\log \frac{1}{2} - \Delta(\delta, \alpha, d, k) \right] \\
&= \Delta(\delta, \alpha, d, k)^2,
\end{aligned}$$

where the inequality follows from the concavity of $p \mapsto \log(p)$.

B.5 Proof of Lemma 9

Recall that for any two probability distributions P, Q , Pinsker's inequality [8] asserts that the total variation norm is bounded as $\|P - Q\|_{\text{TV}} \leq \sqrt{D_{\text{kl}}(P \| Q) / 2}$. Applying this inequality in our setting, we find that

$$\|Q_1^n - Q_{-1}^n\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{kl}}(Q_1^n \| Q_{-1}^n)} = \frac{1}{\sqrt{2}} \sqrt{n D_{\text{kl}}(Q_1 \| Q_{-1})},$$

where we have exploited the product nature of Q_i^n . Now we note that by the concavity of the log, we have (via the first-order inequality) that $\log \frac{1+\delta}{1-\delta} \leq 2\delta/(1-\delta)$, so

$$\frac{1+\delta}{2} \log \frac{\frac{1+\delta}{2}}{\frac{1-\delta}{2}} + \frac{1-\delta}{2} \log \frac{\frac{1-\delta}{2}}{\frac{1+\delta}{2}} = \frac{1+\delta}{2} \log \frac{1+\delta}{1-\delta} + \frac{1-\delta}{2} \log \frac{1-\delta}{1+\delta} = \delta \log \frac{1+\delta}{1-\delta} \leq \frac{2\delta^2}{1-\delta}.$$

Assuming that $\delta \leq 1/3$, the final term is upper bounded by $3\delta^2$. But of course by definition of Q_1 and Q_{-1} , we have

$$D_{\text{kl}}(Q_1 \| Q_{-1}) = \frac{1+\delta}{2} \log \frac{\frac{1+\delta}{2}}{\frac{1-\delta}{2}} + \frac{1-\delta}{2} \log \frac{\frac{1-\delta}{2}}{\frac{1+\delta}{2}} \leq 3\delta^2,$$

which completes the proof.

C Background on Conditional Probabilities

In this appendix, we present some basic lemmas on conditional independence and regular conditional probabilities that will be useful in Appendix D.

We first recall the following classical data-processing inequality, which holds for essentially arbitrary random variables [21, Chapter 5]:

Lemma 11 (Data processing). *Let $X \rightarrow Z \rightarrow Y$ be a Markov chain. Then $I(X; Y) \leq I(X; Z)$, with equality if and only if X is conditionally independent of Y given Z .*

This inequality, in conjunction with with Carathéodory and Minkowski's finite-dimensional version of the Krein-Milman theorem [e.g. 24], allows us to argue any Q minimizing $I(P, Q)$ must be supported on the extreme points of D . To make this point precise, however, we need to address certain measurability issues involved in the choice of the extreme points.

We begin with a precise definition of a regular conditional probability.

Definition 5. Let (Ω, \mathcal{F}) and $(T, \sigma(T))$ be measurable spaces. A *regular conditional probability*, also known as a Markov kernel or transition probability, is a function $\nu : T \times \mathcal{F} \rightarrow [0, 1]$ such that

$$\begin{aligned} t \mapsto \nu(t, A) & \text{ is measurable for all } A \in \mathcal{F} \\ \nu(t, \cdot) : \mathcal{F} & \rightarrow [0, 1] \text{ is a probability measure for all } t \in T. \end{aligned}$$

Any Markov chain has a transition probability; conversely, any set of consistent transition probabilities define a Markov chain (see, e.g., Chapter 5 of Kallenberg [25]).

Some difficulties with measurability arise in constructing the appropriate Markov chain for our setting. To deal with them, we use results from Choquet theory, which extend Krein-Milman theorems to integral representations [35]. We begin our proof by stating a measurable selection theorem [35, Theorem 11.4], though we restrict the theorem's statement to subsets of finite dimensional space.

Proposition 4. *Let $D \subset \mathbb{R}^d$ be a compact convex set. For each x , there exists a probability measure μ_x supported on $\text{Ext}(D)$ such that $\int_D y d\mu_x(y) = x$. Moreover, the mapping $x \mapsto \mu_x$ can be taken to be measurable.*

In the statement of this result, measurability is taken with respect to the σ -field generated by the topology of weak convergence. As a consequence of the proposition, however, it is clear that since for any continuous function f the mapping $x \mapsto \int f d\mu_x$ is measurable, we have that for relatively open sets $A \subset C$ the mapping $x \mapsto \mu_x(A)$ is measurable, whence for any measurable set $A \subset C$ the mapping $x \mapsto \mu_x(A)$ is measurable. That is, we can define the Markov kernel $\nu : \mathbb{R}^d \times \sigma(C) \rightarrow [0, 1]$ according to the mapping specified by Proposition 4 (we take $\nu(x, \cdot) = \mu_x$) with the additional properties that

$$\int_D y \nu(x, dy) = x \quad \text{and} \quad \nu(x, D \setminus \text{Ext}(D)) = 0 \quad \text{for all } x \in D.$$

In finite dimensions, a trivial extension of Proposition 4 allows us to drop the assumption that D is convex. Indeed, we have that since D is compact, then $\text{Ext}(D) = \text{Ext}(\text{Conv}(D))$ [24, Chapter III.2].

Given this measure-theoretic background, we turn to a key lemma that we will need in Appendix D. In this lemma, we assume as usual that $C \subset D \subset \mathbb{R}^d$ are compact sets, and that $Q \in \mathcal{Q}(C, D)$ (recall the definition (9b)).

Lemma 12. *Let P be a distribution supported on C . If there exists a set $A \subset C$ with $P(A) > 0$ and a set $B \subset D \setminus \text{Ext}(D)$ with $Q(B \mid X = x) > 0$ for $x \in A$, there exists a regular conditional probability distribution $Q' \in \mathcal{Q}(C, D)$ where $Q'(\cdot \mid x)$ has support contained in $\text{Ext}(D)$ and*

$$I(P, Q) > I(P, Q').$$

Paraphrasing the lemma slightly, we have that *any* conditional distribution Q minimizing $I(P, Q)$ must (outside of a set of measure zero) be completely supported on the extreme points $\text{Ext}(D)$.

Proof For any $y \in D$, Proposition 4 guarantees that we can represent y as the (regular conditional) measure $\nu(y, \cdot)$. Thus we can define a random variable Z_y distributed according to $\nu(y, \cdot)$, whose existence we are guaranteed by standard constructions [4, 25] with regular conditional probability. Then $\mathbb{E}[Z_y] = \int_D z \nu(y, dz) = y$, and moreover, we can define the measurable version of the conditional expectation $\mathbb{E}[Z_Y | Y]$ via

$$\mathbb{E}[Z_Y | Y] = \int_D z \nu(Y, dz) = Y$$

so we have the (almost sure) chain of equalities

$$\begin{aligned} \mathbb{E}[Z_Y | X = x] &= \mathbb{E}[\mathbb{E}[Z_Y | Y] | X = x] = \int_D \mathbb{E}[Z_Y | Y = y] dQ(y | X = x) \\ &= \int_D \int_D z \nu(y, dz) dQ(y | X = x) = \int_D y dQ(y | X = x) = x. \end{aligned}$$

By construction, $X \rightarrow Y \rightarrow Z$ is a valid Markov chain, and since the sets A and B satisfy $P(A) > 0$ and $\int_A Q(B | X = x) dP(x) > 0$, we see that $I(X; Y) > I(X; Z)$ by Lemma 11. \square

We turn to an analogue of Lemma 12 in the differentially private setting.

Lemma 13. *Let the conditions of Lemma 12 hold, and let P be a distribution supported on C . If there exists a set $A \subset C$ with $P(A) > 0$ and a set $B \subset D \setminus \text{Ext}(D)$ with $Q(B | X = x) > 0$ for $x \in A$, there exists a regular conditional probability distribution $Q' \in \mathcal{Q}(C, D)$ where $Q'(\cdot | x)$ has support contained in $\text{Ext}(D)$, satisfies*

$$I(P, Q) > I(P, Q'),$$

and has no worse differential privacy than Q :

$$\sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q'(S | X = x)}{Q'(S | X = x')} \leq \sup_{S \in \sigma(D)} \sup_{x, x' \in C} \frac{Q(S | X = x)}{Q(S | X = x')}.$$

Proof Let $\nu : \mathbb{R}^d \times \sigma(C) \rightarrow [0, 1]$ be the Markov kernel defined in the proof of Lemma 12, and without loss of generality assume that $Q(\cdot | X = x)$ and $Q(\cdot | X = x')$ have density q with respect to an underlying measure $\mu_{x, x'}$. Define the distribution

$$Q'(S | X = x) := \int_D \int_D \nu(y, dz) q(y | x) d\mu_{x, x'}(y).$$

By assumption, if Q is α -differentially private, then for μ -almost all $y \in D$, we have $q(y | x) \leq e^\alpha q(y | x')$. We find that

$$\begin{aligned} Q'(S | X = x) &= \int_D \int_D \nu(y, dz) q(y | x) d\mu_{x, x'}(y) \\ &\leq \int_D \int_D \nu(y, dz) e^\alpha q(y | x') d\mu_{x, x'}(y) = e^\alpha Q'(S | X = x'), \end{aligned}$$

so Q' is at least as differentially private as Q . \square

Finally, we will need the following standard maximum entropy result. Let z denote a discrete random variable and let $q(z | x)$ denote the conditional probability mass function of $Z | X = x$. Consider the finite dimensional entropy maximization problem

$$\begin{aligned} & \underset{q}{\text{minimize}} \quad \sum_z q(z | x) \log q(z | x) \\ & \text{subject to} \quad \sum_z z q(z | x) = x, \quad \sum_z q(z | x) = 1, \quad q(z | x) \geq 0 \text{ for all } z. \end{aligned} \quad (42)$$

We have the following lemma, which establishes the form of the solution to the problem (42). We include a proof for completeness.

Lemma 14. *The p.m.f. $q(\cdot | x)$ solving problem (42) is given by*

$$q(z | x) = \frac{\exp(-\mu^\top z)}{\sum_{z'} \exp(-\mu^\top z')}, \quad (43)$$

where $\mu \in \mathbb{R}^d$ is any vector chosen to satisfy the constraint $\sum_z z q(z | x) = x$. Such a $\mu \in \mathbb{R}^d$ exists.

Proof We may write the Lagrangian with dual variables $\mu \in \mathbb{R}^d$, $\lambda(z) \geq 0$, and $\theta \in \mathbb{R}$,

$$\mathcal{L}(q, \mu, \lambda, \theta) = \sum_z q(z | x) \log q(z | x) + \mu^\top \left(\sum_z z q(z | x) - x \right) + \theta \left(\sum_z q(z | x) - 1 \right) - \sum_z \lambda(z) q(z | x).$$

Since the problem (42) has convex cost, linear constraints, and non-empty domain, strong duality obtains [6, Chapter 5], and the KKT conditions hold for the problem. Thus, minimizing q out of \mathcal{L} to find the dual, we take derivatives with respect to the m variables $q(z | x)$ for $z = (1 + \alpha)u_i$ and find the optimal conditional p.m.f. q must satisfy

$$\log q(z | x) + 1 + \mu^\top z + \theta - \lambda(z) = 0, \quad \text{or} \quad q(z | x) = \exp(\lambda(z) - 1 - \theta) \exp(-\mu^\top z).$$

In particular, we see that since $q(z | x) > 0$, we must have $\lambda(z) = 0$ by complementarity, and (satisfying the summability constraint $\sum_z q(z | x) = 1$) we see that

$$q(z | x) = \frac{\exp(-\mu^\top z)}{\sum_{z'} \exp(-\mu^\top z')},$$

where $\mu \in \mathbb{R}^d$ is any vector chosen to satisfy the constraint $\sum_z z q(z | x) = x$. The existence of such a μ is guaranteed by the attainment of the KKT conditions. \square

D Proofs of Minimax Mutual Information Characterizations

In this section, we provide the proofs of the results stated in Section 4, all of which follow a broadly similar outline. We make use of Lemma 12 to guarantee that any conditional distribution Q minimizing the mutual information $I(P, Q)$ must be supported on the extreme points of the set D . This allows us to reduce computing maximal entropies and minimal mutual information values to finite dimensional convex programs, whose optimality we can check using results from convex analysis and optimization.

D.1 Proof of Theorem 4

We begin by considering \sup_P , where Q^* is defined as in the statement of the theorem. Since the support of Q^* is finite (there are m extreme points of D), we have

$$\begin{aligned} I(P, Q^*) &= I(X; Z) = H(Z) - H(Z | X) \leq \log(m) - H(Z | X) \\ &= \log(m) - \int H(Z | X = x) dP(x). \end{aligned}$$

Now, for any distribution P on the set C and for any $x \in \text{supp } P$, we can write x as $x = \sum_i \lambda_i(x) u_i$, where u_i are the extreme points of C , and where $\lambda_i(x) \geq 0$ and $\sum_i \lambda_i(x) = 1$ (using the Krein-Milman theorem). Define the individual probability mass functions q^i to be the maximum entropy p.m.f. (43) for each of the extreme points u_i . Then we can define the conditional probability mass function by

$$q(\cdot | x) = \sum_i \lambda_i(x) q^i(\cdot).$$

(Without loss of generality, we may assume that the λ_i are continuous, since the set of extreme points is finite, and thus $q(\cdot | x)$ can be viewed as a regular conditional probability. We can make this formal using the techniques in the proof of Lemma 12.) Denoting $H(q(\cdot | x)) := H(Z | X = x)$, we can use the convexity of the negative entropy to see that

$$I(P, Q^*) \leq \log(m) - \int \sum_i \lambda_i(x) H(q^i(\cdot)) dP(x). \quad (44)$$

By symmetry, the entropy $H(q^i(\cdot)) = H(Q^*(\cdot | X = u_i))$ is a constant determined by the maximum entropy distribution (43), and thus

$$I(P, Q^*) \leq \log(m) - H(Q^*(\cdot | X = u_i)). \quad (45)$$

Equality in the upper bound (45) is attained by taking P^* to be the uniform distribution on the extreme points $\{u_i\}$ of C .

It remains to establish an identical lower bound for $I(P^*, Q)$ over all conditional distributions Q satisfying the constraints of the theorem statement. We know from Lemma 12 that Q must be supported on $(1 + \kappa)u_i$ for $i = 1, \dots, m$. Denoting by $q(z | x)$ the p.m.f. of Q conditional on x (for x in the finite set of extreme points of C that make up the support $\text{supp } P^*$), we can write minimizing the mutual information as the parametric convex optimization problem

$$\begin{aligned} &\underset{q}{\text{minimize}} \quad \sum_z \left(\sum_x q(z | x) p(x) \right) \log \left(\sum_x q(z | x) p(x) \right) - \sum_x p(x) \sum_z p(z | x) \log p(z | x) \quad (46) \\ &\text{subject to} \quad \sum_z p(z | x) = 1 \text{ for all } x, \quad \sum_z z p(z | x) = x \text{ for all } x, \quad p(z | x) \geq 0 \text{ for all } x, z. \end{aligned}$$

In the problem (46), the sums over x and z are over the extreme points of C and D , respectively and p is the uniform distribution with $p(x) = 1/m$. Mutual information is convex in the conditional distribution q ; moreover, it is strictly convex except when $q(z | x) = \sum_{x'} q(z | x') p(x')$ for all x, z . (This can be seen by an inspection of the proof of Theorem 2.7.4 by Cover and Thomas [8].) In our case, since Q^* does not satisfy this equality, the uniqueness of Q^* as the minimizer of $I(P^*, Q^*)$ will follow if we show that Q^* is a minimizer at all.

We proceed to solve the problem (46). Writing $I(p, q)$ as a shorthand for the mutual information, we introduce Lagrange multipliers $\theta(x) \in \mathbb{R}$ for the normalization constraints, $\mu(x) \in \mathbb{R}^d$ for the conditional expectation constraints, and $\lambda(x, z) \geq 0$ for the nonnegativity constraints. This yields the Lagrangian

$$\mathcal{L}(q, \mu, \lambda, \theta) = I(p, q) - \sum_{x, z} \lambda(x, z) q(z | x) + \sum_x \mu(x)^\top \left(\sum_z z q(z | x) - x \right) + \sum_x \theta(x) \left(\sum_z q(z | x) - 1 \right).$$

If we can satisfy the Karush-Kuhn-Tucker (KKT) conditions (see, e.g., [6]) for optimality of the problem (46), we will be done. Taking derivatives with respect to $q(z | x)$, we see

$$\begin{aligned} \frac{\partial}{\partial q(z | x)} \mathcal{L}(q, \mu, \lambda, \theta) &= p(x) [\log(q(z | x)) + 1] - p(x) \log \left(\sum_{x'} q(z | x') p(x') \right) \\ &\quad - q(z) \cdot \frac{1}{q(z)} p(x) - \lambda(z, x) + \theta(x) + \mu(x)^\top z \\ &= p(x) \log q(z | x) - p(x) \log \left(\sum_{x'} q(z | x') p(x') \right) - \lambda(z, x) + \theta(x) + \mu(x)^\top z, \end{aligned}$$

where we set $q(z) = \sum_{x'} q(z | x') p(x')$ for shorthand. Now, we use symmetry to note that since we have chosen q to be the maximum entropy distribution (43) for each x in the extreme points $\{u_i\}$ of C , the marginal $q(z) = \sum_{x'} q(z | x') p(x') = 1/m$ is uniform by the symmetry of the set D and since p is uniform. In addition, since $q(z | x) > 0$ strictly, we have $\lambda(z, x) = 0$ by complementarity. Thus, at q chosen to be the maximum entropy distribution, we can rewrite the derivative of the Lagrangian

$$\frac{\partial}{\partial q(z | x)} \mathcal{L}(q, \mu, \lambda, \theta) = \frac{1}{m} \log q(z | x) - \frac{1}{m} \log \frac{1}{m} + \theta(x) + \mu(x)^\top z.$$

Recalling the definition (43) of $q(z | x)$, and denoting the maximum entropy parameters μ there by $\mu^*(x)$, we have

$$\frac{\partial}{\partial q(z | x)} \mathcal{L}(q, \mu, \lambda, \theta) = -\frac{1}{m} \mu^*(x)^\top z + \frac{1}{m} \log \left(\sum_{z'} \exp(-\mu^*(x)^\top z') \right) - \frac{1}{m} \log \frac{1}{m} + \theta(x) + \mu(x)^\top z.$$

Now, by inspection we may set

$$\theta(x) = \frac{1}{m} \log \frac{1}{m} - \frac{1}{m} \log \left(\sum_{z'} \exp(-\mu^*(x)^\top z') \right) \quad \text{and} \quad \mu(x) = \frac{1}{m} \mu^*(x),$$

and we satisfy the KKT conditions for the mutual information minimization problem (46).

Summarizing, the conditional distribution Q^* specified in the statement of the theorem as the maximum entropy distribution (43) satisfies

$$\inf_Q I(P^*, Q) \geq I(P^*, Q^*),$$

which, when combined with the first part of the proof, gives the saddle point inequality

$$\sup_P I(P, Q^*) \leq \log(m) - H(q(\cdot | X = u_i)) = I(P^*, Q^*) \leq \inf_Q I(P^*, Q),$$

as claimed.

Remarks: In the proof of the theorem, we have defined $Q^*(\cdot | x)$ as a conditional distribution only for $x \in \text{Ext}(C)$, the extreme points of C . This can easily be remedied: take $Q^*(\cdot | x)$ to be the distribution maximizing the entropy $H(Z | X = x)$ for each $x \in C$ under the constraint that the support of Z be contained in $\text{Ext}(D)$. This is equivalent to—for each $x \in C$ —choosing $Z = z_i$ for $z_i \in \text{Ext}(D)$, $i = 1, \dots, m$, with probability q_i , where $q \in \mathbb{R}^m$ solves the entropy maximization problem

$$\underset{q \in \mathbb{R}^m}{\text{maximize}} \quad - \sum_i q_i \log q_i \quad \text{subject to} \quad \sum_i z_i q_i = x, \quad \sum_i q_i = 1, \quad q_i \geq 0.$$

Inspecting the proof of Theorem 4 (see the bound (44)) shows that this choice can only decrease the mutual information $I(X; Z)$. Additionally, the strong convexity of the entropy over the simplex guarantees that the solutions to this optimization problem are continuous in x (see Chapter X of Hiriart-Urruty and Lemaréchal [24]) so this distribution $q(\cdot | x)$ defines a measurable random variable as desired.

Additionally, though Theorem 4 assumes that the sets C and D satisfy $D = (1 + \kappa)C$ for some $\kappa > 0$, inspection of the proof yields a somewhat stronger result. Assume the distribution Q maximizing the entropy $H(Z | X = x)$ satisfies $H(Q(\cdot | X = x)) = H(Q(\cdot | X = x'))$ for each extreme point x of C and additionally satisfies that for each extreme point z of D the sum $\sum_x Q(Z = z | X = x)$ is a constant (the sum is over extreme points x of C). Then the upper bound (45) is attained with equality, and a similar calculation yields that Q solves the mutual information problem (46). Thus, as long as C and D are suitably jointly symmetric, Z should be chosen to maximize the entropy $H(Z | X = x)$ for each $x \in C$.

D.2 Proof of Proposition 1

Using Theorem 4 (and the remarks immediately following its proof), we can focus on maximizing the entropy of the random variable Z conditional on $X = x$ for each fixed $x \in [-1, 1]^d$. Let Z_i denote the i th coordinate of the random vector Z ; we take the conditional distribution of Z_i to be independent of Z_j and let Z be distributed as

$$Z_i | X = \begin{cases} M & \text{w.p. } \frac{1}{2} + \frac{X_i}{2M} \\ -M & \text{w.p. } \frac{1}{2} - \frac{X_i}{2M}. \end{cases} \quad (47)$$

Let us now verify that the distribution (47) maximizes the entropy $H(Z | X = x)$. Indeed, ignoring the conditioning we write the entropy maximization problem

$$\underset{q}{\text{minimize}} \quad -H(q) \quad \text{subject to} \quad \sum_z q(z) = 1, \quad q(z) \geq 0, \quad \sum_z z q(z) = x, \quad (48)$$

where all sums are taken over $z \in \text{Ext}([-M, M]^d) = \{-M, M\}^d$. Introducing the Lagrange multipliers $\mu \in \mathbb{R}^d$, $\lambda(z) \geq 0$, and $\theta \in \mathbb{R}$, we find that problem (48) has the Lagrangian

$$\mathcal{L}(q, \mu, \lambda, \theta) = -H(q) - \sum_z \lambda(z) q(z) + \mu^\top \left(\sum_z z q(z) - x \right) + \theta \left(\sum_z q(z) - 1 \right).$$

To find the infimum of the Lagrangian with respect to q , we take derivatives (since we make the identification $q \in \mathbb{R}^{2^d}$). We see that

$$\frac{\partial}{\partial q(z)} \mathcal{L}(q, \mu, \lambda, \theta) = \log(q(z)) + 1 - \lambda(z) + \theta + \mu^\top z.$$

With the definition (47) of the probability mass function q (that z_i are independent Bernoulli random variables with parameters $\frac{1}{2} + x_i/2M$), the coordinate conditional distributions are

$$q(z_i | x_i) = \left(\frac{1}{2} + \frac{1}{2M}\right)^{\frac{1}{2} + \frac{x_i z_i}{2M}} \left(\frac{1}{2} - \frac{1}{2M}\right)^{\frac{1}{2} - \frac{x_i z_i}{2M}}.$$

Theorem 4 says that without loss of generality we may assume that $x \in \{-1, 1\}^d$, the full probability mass function q can be written

$$q(z) = \left(\frac{1}{2} + \frac{1}{2M}\right)^{\frac{d}{2} + \frac{x^\top z}{2M}} \left(\frac{1}{2} - \frac{1}{2M}\right)^{\frac{d}{2} - \frac{x^\top z}{2M}}. \quad (49)$$

Plugging the conditional (49) results in

$$\begin{aligned} & \frac{\partial}{\partial q(z)} \mathcal{L}(q, \mu, \lambda, \theta) \\ &= \left(\frac{d}{2} + \frac{x^\top z}{2M}\right) \log \left(\frac{1}{2} + \frac{1}{2M}\right) + \left(\frac{d}{2} - \frac{x^\top z}{2M}\right) \log \left(\frac{1}{2} - \frac{1}{2M}\right) + 1 - \lambda(z) + \theta + \mu^\top z \\ &= \frac{d}{2} \left[\log \left(\frac{1}{2} + \frac{1}{2M}\right) + \log \left(\frac{1}{2} - \frac{1}{2M}\right) \right] + \frac{x^\top z}{2M} \left[\log \left(\frac{1}{2} + \frac{1}{2M}\right) - \log \left(\frac{1}{2} - \frac{1}{2M}\right) \right] \\ & \quad + 1 - \lambda(z) + \theta + \mu^\top z. \end{aligned}$$

Performing a few algebraic manipulations with the logarithmic terms, the final equality becomes

$$d \log \left(\frac{\sqrt{(M+1)(M-1)}}{M} \right) + \frac{x^\top z}{M} \log \left(\sqrt{\frac{M+1}{M-1}} \right) + 1 - \lambda(z) + \theta + \mu^\top z.$$

The complementarity conditions for optimality [6] imply that $\lambda(z) = 0$, and since the equality constraints in the problem (48) are satisfied, we can choose θ and μ arbitrarily. Taking

$$\theta = -d \log \left(\frac{\sqrt{(M+1)(M-1)}}{M} \right) - 1 \quad \text{and} \quad \mu = -x \frac{1}{M} \log \left(\sqrt{\frac{M+1}{M-1}} \right)$$

yields that the partial derivatives of \mathcal{L} are 0, which shows that indeed our choice of Q^* is optimal.

D.3 Proof of Proposition 2

The proof follows along lines similar to the ℓ_∞ case: we compute the maximum entropy distribution subject to the constraint that $\mathbb{E}[Z] = x$ for some $x \in \mathbb{R}^d$ with $\|x\|_1 \leq 1$, and Z must be supported on the extreme points $\pm M e_i$ of the ℓ_1 -ball of radius M . (Recall that $e_i \in \mathbb{R}^d$ are the standard basis vectors.) Based on Theorem 4, in order to find the minimax mutual information, we need only consider the cases where $x = \pm e_i$ for some $i \in \{1, \dots, d\}$.

Following this plan, we recall the entropy maximization problem (48), where now $x = \pm e_i$ and the sums are over $z \in M\{\pm e_i\}_{i=1}^d$. As in the proof of Proposition 1, we can write the Lagrangian and take its derivatives, finding that for $z = \pm M e_i$ we have

$$\frac{\partial}{\partial q(z)} \mathcal{L}(q, \mu, \lambda, \theta) = \log(q(z)) + 1 - \lambda(z) + \theta - \mu^\top z.$$

Solving for $q(z)$, we find that

$$q(z) = \exp(\lambda(z) - 1 - \theta) \exp(\mu^\top z),$$

but complementarity [6] guarantees that $\lambda(z) = 0$ since $q(z) > 0$, and normalizing we may write $q(z) = \exp(-\mu^\top z) / \exp(-\mu^\top \sum_{z'} z')$, where the sum is over the extreme points of the ℓ_1 -ball of radius M . In particular, $q(Me_i) \propto e^{-\mu_i}$ and $q(-Me_i) \propto e^{\mu_i}$. Without loss of generality, let $x = e_i$. Symmetry suggests we take (and we verify this to be true)

$$q(z) = \exp(-1 - \theta) \begin{cases} \exp(\mu_i) & \text{if } z = Me_i \\ \exp(-\mu_i) & \text{if } z = -Me_i \\ \exp(0) & \text{otherwise.} \end{cases} \quad (50)$$

Indeed, with the choice (50) of q , we have $q(Me_j) - q(-Me_j) = 0$ for $j \neq i$, while (setting $\gamma = \mu_i$ and normalizing appropriately)

$$q(Me_i) - q(-Me_i) = \frac{e^\gamma}{e^{-\gamma} + e^\gamma + 2(d-1)} - \frac{e^{-\gamma}}{e^{-\gamma} + e^\gamma + 2(d-1)}.$$

Thus, if we can solve the equation $Mq(Me_i) - Mq(-Me_i) = 1$, we will be nearly done. To that end, we write

$$\frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2(d-1)} = \frac{1}{M} \quad \text{or} \quad \beta - \beta^{-1} = \frac{1}{M} (\beta + \beta^{-1} + 2(d-1)),$$

where we identified $\beta = e^\gamma$. Multiplying both sides by β , we have a quadratic equation in β :

$$\beta^2 - 1 = \frac{1}{M} (\beta^2 + 2\beta(d-1) + 1) \quad \text{or} \quad (M-1)\beta^2 - 2(d-1)\beta - (M+1) = 0,$$

whose solution is the positive root of

$$\beta = \frac{2d-2 \pm \sqrt{(2d-2)^2 + 4(M^2-1)}}{2(M-1)} \quad \text{or} \quad \gamma = \log \left(\frac{2d-2 + \sqrt{(2d-2)^2 + 4(M^2-1)}}{2(M-1)} \right).$$

By our construction, with γ so defined, we satisfy the constraints that $M[q(Me_i) - q(-Me_i)] = 1$ and $q(Me_j) - q(-Me_j) = 0$ for $j \neq i$. Since q belongs to the exponential family and satisfies the constraints, it maximizes the entropy $H(Z)$ as desired [8].

Algebraic manipulations and the computation of the conditional entropy $H(Z | X = e_i)$ give the remainder of the statement of the proposition.

D.4 Proof of Proposition 3

The outline of the proof of Proposition 3 is as follows. First, recall from Lemma 13 that any distribution satisfying optimal local differential privacy must be supported on the extreme points of the outer set D (as in the proof of Theorem 4). Given this result, we reduce the problem of finding an optimally private distribution to a linear program, using symmetry arguments to simplify the LP. Finally, we show that the solution to the linear program is unique, which means that we have found the unique distribution satisfying optimal local differential privacy.

We begin by developing a reduction of the problem of finding a distribution with optimal local differential privacy to a linear program. Note that there is a non-increasing mapping between M —the radius of the larger ℓ_∞ ball—and α^* . Indeed, whenever M increases, the set of distributions Q from which to choose a privacy channel increases, so α^* decreases. Put inversely, for a given differential privacy level α , we can find the smallest M such that it is possible to construct an α -differentially private channel Q mapping from $[-1, 1]^d$ to $[-M, M]^d$. (Lemma 16 shows that the mapping from M to α^* is implicitly invertible.)

We thus take the view of finding the largest M such that an α -differentially private distribution exists. Fix $d \in \mathbb{N}$ and (with some abuse of notation) let $Z \in \{-1, 1\}^{d \times 2^d}$ be the matrix whose columns are the edges of the hypercube $\{-1, 1\}^d$. For each $z, x \in \{-1, 1\}^d$, define the variables $q(z | x) \geq 0$ to represent the conditional probabilities, using $q(x) \in [0, 1]^{2^d}$ to represent the vector $[q(z | x)]_{z \in \{-1, 1\}^d}$. Then the linear program

$$\begin{aligned} & \text{minimize} && -t \\ & \text{subject to} && Zq(x) - tx = 0 \text{ for all } x \in \{-1, 1\}^d \\ & && q(z | x) \leq e^\alpha q(z | x') \text{ for all } x, x', z \in \{-1, 1\}^d \\ & && \sum_z q(z | x) = 1, \quad q(x) \succeq 0 \text{ for all } x \in \{-1, 1\}^d, \end{aligned} \tag{51}$$

at its solution t^* (this solution is guaranteed to exist, since the vectors q live in a compact set), yields the smallest value $M = 1/t^*$ for which it is possible to have an α -differentially private channel Q . The solution vectors $q(x)$ give one such channel.

It is possible to calculate the solution of the LP (51) by hand, but it is tedious. We thus use the structure of optimal local differential privacy to reduce the problem to a single minimization problem over a vector $q \in \mathbb{R}^{2^d}$ (rather than a matrix $[q(x)] \in \mathbb{R}^{2^d \times 2^d}$). We have

Lemma 15. *A distribution satisfying optimal local differential privacy must, for each $x \in \{-1, 1\}^d$, have $q(x) = \Pi(x)q$, where $\Pi(x) \in \{0, 1\}^{2^d \times 2^d}$ is a permutation matrix and q is a fixed vector.*

Proof Suppose for the sake of contradiction that this is not the case, but the vectors $q(x)$ and t solve the linear program (51). Let Q_1 denote the matrix of the vectors $q(x)$. Choose vectors $q(x)$ and $q(x')$ such that $q(x) \neq \Pi q(x')$ for any permutation matrix Π . Now construct vectors $q_2(x)$ and $q_2(x')$ such that $q_2(z | x) = q(z' | x')$, where z' is chosen so that $z'_i x'_i = z_i x_i$, and similarly choose q_2 so that $q_2(z | x') = q(z' | x)$, where $z_i x_i = z'_i x'_i$. Let Q_2 denote the matrix of vectors q , but where $q_2(x)$ and $q_2(x')$ replace $q(x), q(x')$. Then by construction, all the constraints of the original linear program (51) are satisfied. By symmetry and the strict convexity of the mutual information in the channel distribution Q , however, we see that

$$I(P, Q_1) = I(P, Q_2) = \frac{1}{2} (I(P, Q_1) + I(P, Q_2)) > I\left(P, \frac{1}{2}(Q_1 + Q_2)\right).$$

The decrease in mutual information gives the necessary contradiction. \square

With Lemma 15 in hand, we can now turn to the smaller linear program—in a single vector q and for a single vector $x \in \{-1, 1\}^d$ —that will give us the locally optimal differentially private

channel. Indeed, we consider the linear program in the variables $t \in \mathbb{R}$ and $q \in \mathbb{R}^{2^d}$, where q is indexed by the column z of Z .

$$\begin{aligned} & \text{minimize} && -t \\ & \text{subject to} && Zq - tx = 0 \\ & && q(z) \leq e^\alpha q(z') \text{ for all } z, z', \\ & && \sum_z q(z) = 1, \quad q \geq 0. \end{aligned} \tag{52}$$

Define the constants

$$K_d = \sum_{i=0}^{\lfloor d/2 \rfloor} (d-2i) \binom{d}{i} \quad \text{and} \quad C_d = \text{card} \left\{ z \in \{-1, 1\}^d : z^\top x > 0 \right\} = \begin{cases} 2^{d-1} & \text{if } d \text{ odd} \\ 2^{d-1} - \frac{1}{2} \binom{d}{d/2} & \text{if } d \text{ even.} \end{cases}$$

We have the following lemma, which characterizes the structure of the solution vector q .

Lemma 16. *Define $\alpha^* = \log \frac{K_d + 2^d - C_d}{K_d - C_d}$. For any $\alpha < \alpha^*$, the unique solution to the linear program (52) is given by*

$$q(z) = \begin{cases} \frac{e^\alpha}{e^\alpha C_d + 2^d - C_d} & \text{if } \langle z, x \rangle > 0 \\ \frac{1}{e^\alpha C_d + 2^d - C_d} & \text{otherwise.} \end{cases}$$

Proof First, problem (52) is clearly equivalent to the linear program

$$\begin{aligned} & \text{minimize} && -t \\ & \text{subject to} && Zq - tx = 0 \\ & && \max_z \{q(z)\} + e^\alpha \max_z \{-q(z)\} \leq 0 \\ & && \sum_z q(z) = 1, \quad q \geq 0. \end{aligned} \tag{53}$$

Our proof proceeds in two large steps: first, we argue that a q of the form specified in the lemma is indeed the solution to the problem (53), then we use results on uniqueness of solutions to linear programs due to Mangasarian [32].

For the first step, we begin by writing the Lagrangian to the problem (53). We introduce dual variables $\theta \in \mathbb{R}^{2^d}$ for the constraint $Zq - tx = 0$, $\lambda \geq 0$ for the first inequality, $\tau \in \mathbb{R}$ for the sum constraint, and $\beta \in \mathbb{R}_+^{2^d}$ for the non-negativity of q . With this, we have Lagrangian

$$\mathcal{L}(q, t, \theta, \lambda, \tau, \beta) = -t + \theta^\top \left(\sum_z q(z) - tx \right) + \lambda \max_z \{q(z)\} + e^\alpha \max_z \{-q(z)\} + \tau(\mathbb{1}^\top q - 1) - \beta^\top q. \tag{54}$$

Recall the generalized subgradient KKT conditions for optimality of the solution to an optimization problem [24, Chapter VII]. A vector $q > 0$ is optimal for the problem (53) if the constraints $\max_i \{q_i\} \leq e^\alpha \min_i \{q_i\}$ and $\sum_i q_i = 1$ hold, there is a $t \geq 0$ such that $Zq - tx = 0$, and we can find θ , λ , and τ such that

$$Z^\top \theta + \lambda [v_+ - e^\alpha v_-] + \tau \mathbb{1} = 0, \quad \beta = 0, \quad \text{and} \quad \theta^\top x = -1, \tag{55}$$

where v_+ and v_- are vectors satisfying

$$v_+ \in \text{Conv} \left\{ e_i : q_i = \max_i \{q_i\} \right\} \quad \text{and} \quad v_- \in \text{Conv} \left\{ e_i : q_i = \min_i \{q_i\} \right\}.$$

That $\beta = 0$ follows by complementarity (recall that $q > 0$ is assumed).

If we can find settings for the vectors θ, λ, τ , and v_{\pm} satisfying the KKT conditions (55), we are done. To that end, set $\theta = -x/d$. Then by inspection $\theta^\top x = -\|x\|_2^2/d = -1$, and we can rewrite the remaining KKT condition by noting that we must find vectors v_+, v_- , and $\tau \in \mathbb{R}$ such that

$$\begin{aligned} -\frac{1}{d}Z^\top x + v_+ - e^\alpha v_- + \tau \mathbf{1} &= 0, \quad v_+^\top \mathbf{1} = v_-^\top \mathbf{1}, \quad v_+ \geq 0, v_- \geq 0, \\ v_+(z) &= 0 \text{ if } q(z) < \max_z \{q(z)\}, \quad \text{and} \quad v_-(z) = 0 \text{ if } q(z) > \min_z \{q(z)\}. \end{aligned}$$

Note that we have eliminated λ as it is a non-negative homogeneous scaling term on v_+ and v_- . Now we assume that we choose the two values q_+, q_- with $0 < q_- < q_+$ such that $q(z) = q_+$ when $z^\top x > 0$ and $q(z) = q_-$ when $z^\top x \leq 0$. (Clearly such values can be chosen such that $\sum_z q(z) = 1$.) We will choose the values of v_+, v_- , and τ satisfying the KKT conditions. Indeed, set

$$v_+(z) = \begin{cases} \frac{z^\top x}{d} - \tau & \text{if } z^\top x > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad v_-(z) = \begin{cases} -e^{-\alpha} \frac{z^\top x}{d} + e^{-\alpha} \tau & \text{if } z^\top x \leq 0 \\ 0 & \text{otherwise} \end{cases} \quad (56)$$

By inspection, we see that $-Z^\top x/d + v_+ - e^\alpha v_- + \tau = 0$, so the only question remaining is whether we can choose τ such that $v_{\pm} \geq 0$ and $v_+^\top \mathbf{1} = v_-^\top \mathbf{1}$.

To that end, we recall the definition of the constant K_d , and we seek τ such that

$$\sum_z v_+(z) = \frac{1}{d}K_d - \tau C_d = e^{-\alpha} \frac{1}{d}K_d + e^{-\alpha} \tau (2^d - C_d) = \sum_z v_-(z)$$

by the symmetry in the sums. Rewriting the equation, we find that for equality we must have

$$\tau \left(e^{-\alpha} (2^d - C_d) + C_d \right) = \frac{1}{d}K_d (1 - e^{-\alpha}) \quad \text{or} \quad \tau = \frac{K_d}{d} \cdot \frac{e^\alpha - 1}{e^\alpha C_d + 2^d - C_d} = \frac{K_d}{dC_d} \cdot \frac{e^\alpha - 1}{e^\alpha + 2^d/C_d - 1}.$$

Thus we find that if α is such that

$$\frac{K_d}{dC_d} \frac{e^\alpha - 1}{e^\alpha + 2^d/C_d - 1} < \frac{1}{d}, \quad (57)$$

then by our choice (56) of the vectors v_+ and v_- , we have $v_+(z) > 0$ whenever $z^\top x > 0$, and $v_-(z) > 0$ whenever $z^\top x \leq 0$. Noting that by our setting of $q(z)$, we have by symmetry of Z that there exists a $t > 0$ such that $Zq = tx$, we find that our choice of q is optimal.

We have two arguments remaining in the proof. The first is to show that for $\alpha < \alpha^*$ defined in the statement of the lemma, the inequality (57) holds. Rewriting the inequality, we solve

$$e^\alpha - 1 = \frac{C_d}{K_d} \left(e^\alpha + 2^d/C_d - 1 \right) \quad \text{or} \quad e^\alpha \left(1 - \frac{C_d}{K_d} \right) = \frac{2^d - C_d}{K_d} + 1, \quad \text{i.e.} \quad \alpha^* = \log \frac{K_d + 2^d - C_d}{K_d - C_d}.$$

For any $\alpha < \alpha^*$, the strict inequality (57) holds, so the setting (56) of v_+ and v_- satisfy the KKT conditions.

Our last argument regards the uniqueness of the two-valued solution vector q . For that, we apply Mangasarian’s result [32, Theorem 1] that if there exists an $\epsilon > 0$ such that for any vector $u \in \mathbb{R}^{2^d}$ with $\|u\|_2 = 1$, q is a solution of the linear program (52) when the objective is $-t + \epsilon u^\top q$, then q is unique. Luckily, this is not difficult given our previous work. The Lagrangian (54) for the modified linear program becomes

$$\epsilon u^\top q - t + \theta^\top \left(\sum_z q(z) - tx \right) + \lambda \max_z \{q(z)\} + e^\alpha \max_z \{-q(z)\} + \tau(\mathbb{1}^\top q - 1) - \beta^\top q.$$

The only modification in our KKT conditions (55) is that the first equality becomes

$$\epsilon u + Z^\top \theta + \lambda[v_+ - e^\alpha v_-] + \tau \mathbb{1} = 0.$$

By the strictness of the inequalities $v_+(z) > 0$ for z such that $z^\top x > 0$ (and similarly for v_-) in the definitions (56) whenever $\alpha < \alpha^*$, we see that for suitably small $\epsilon > 0$, the vectors v_+ and v_- can be perturbed so that the KKT conditions are still satisfied. This proves the uniqueness of the two-valued solution vector q . \square

Remarks: Following an argument with completely the same structure as the proof, we see that for any $d \in \mathbb{N}$ (say $d \geq 3$), there are different “regimes” of α , that is, there exists a sequence $\alpha_0^*, \alpha_2^*, \dots, \alpha_{d-1}^*$ (or α_{d-2}^* if d is even) such that for $\alpha \in (\alpha_{2i}^*, \alpha_{2i+2}^*)$, the unique optimal solution to the linear program (52) is given by taking

$$q(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x \rangle > 2(i+1) \\ 1 & \text{for } z \text{ s.t. } \langle z, x \rangle \leq 2(i+1) \end{cases}$$

(for $\alpha < \alpha_0^*$, we say $i = -1$ above). For $\alpha = \alpha_{2i}^*$, the set of solutions is given by the convex combinations of the solution vectors

$$q_{<}(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x \rangle > 2i \\ 1 & \text{for } z \text{ s.t. } \langle z, x \rangle \leq 2i \end{cases} \quad \text{and} \quad q_{>}(z) \propto \begin{cases} \exp(\alpha) & \text{for } z \text{ s.t. } \langle z, x \rangle > 2(i+1) \\ 1 & \text{for } z \text{ s.t. } \langle z, x \rangle \leq 2(i+1) \end{cases},$$

which follows from arguments similar to our application of Mangasarian’s results [32]. (Recall also the argument with convex combinations of ratios following Proposition 3.)

Now we may complete the proof of Proposition 3. Indeed, we see from Lemma 16 that the distribution satisfying optimal local differential privacy must assign probability masses at two levels—at least when the point being perturbed comes from $\{-1, 1\}^d$. Now let Q be a distribution specified in the lemma. An argument identical to that in our proof of Proposition 1—by symmetry—shows that the distribution P maximizing the mutual information $I(P, Q)$ is uniform on $\{-1, 1\}^d$. The uniqueness of Q then follows from Lemmas 15 and 16, which show that such Q is the only distribution that minimizes the radius M of the ball $[-M, M]^d$; inverting this bound gives the proposition. \square

E Proof of Corollary 4

First, we claim that as $\gamma \rightarrow 0$, the following expansion holds:

$$\log(2d) - \log(e^\gamma + e^{-\gamma} + 2d - 2) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2} = \frac{\gamma^2}{2d} + \Theta\left(\frac{\gamma^4}{d}\right). \quad (58)$$

Before proving this, we use the expansion (58) to prove Corollary 4. Noting that

$$\frac{2d - 2 + \sqrt{(2d - 2)^2 + 4(M^2 - 1)}}{2(M - 1)} = \sqrt{\frac{M + 1}{M - 1}} + \frac{d - 1}{M - 1} + \Theta(d^2/M^2),$$

we see that since $\log(1 + x) = x - x^2/2 + \Theta(x^3)$, we have $\gamma = \frac{d}{M} + \Theta\left(\frac{d^2}{M^2}\right)$. Thus the mutual information in Proposition 2 is

$$\begin{aligned} I(P^*, Q^*) &= \frac{\log^2(\sqrt{(M + 1)/(M - 1)} + d/M + \Theta(d^2/M^2))}{2d} + \Theta\left(\frac{\log^4(1 + d/M)}{d}\right) \\ &= \frac{d}{2M^2} + \Theta\left(\min\left\{\frac{d^3}{M^4}, \frac{\log^4(d)}{d}\right\}\right). \end{aligned}$$

Now we return to showing the claim (58). Indeed, define $f(\gamma) = \log(e^\gamma + e^{-\gamma} + 2d - 2)$. Taking several derivatives, we have

$$f^{(1)}(\gamma) = \frac{e^\gamma - e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2}, \quad f^{(2)}(\gamma) = \frac{(e^\gamma + e^{-\gamma})(2d - 2) + 4}{(e^\gamma + e^{-\gamma} + 2d - 2)^2},$$

and

$$f^{(3)}(\gamma) = \frac{-(e^{2\gamma} - e^{-2\gamma})(2d - 2) - 8(e^\gamma - e^{-\gamma}) + (2d - 2)^2(e^\gamma - e^{-\gamma})}{(e^\gamma + e^{-\gamma} + 2d - 2)^3}.$$

Via a Taylor expansion, we have

$$\log(2d) = \log(e^\gamma - e^{-\gamma} + 2d - 2) + (0 - \gamma)f^{(1)}(\gamma) + \frac{(0 - \gamma)^2}{2}f^{(2)}(\gamma) + \mathcal{O}\left(f^{(3)}(\gamma)\gamma^3\right).$$

Recalling our calculation of the first derivative $f^{(1)}(\gamma)$, we thus see that

$$\begin{aligned} \log(2d) - \log(e^\gamma + e^{-\gamma} + 2d - 2) + \gamma \frac{e^\gamma}{e^\gamma + e^{-\gamma} + 2d - 2} - \gamma \frac{e^{-\gamma}}{e^\gamma + e^{-\gamma} + 2d - 2} \\ = \frac{(e^\gamma + e^{-\gamma})(2d - 2) + 4}{(e^\gamma + e^{-\gamma} + 2d - 2)^2} \cdot \frac{\gamma^2}{2} + \mathcal{O}\left(f^{(3)}(\gamma)\gamma^3\right). \end{aligned}$$

A few simpler Taylor expansions yield that $f^{(3)}(\gamma) = \mathcal{O}(\gamma/d)$, which means that all we have left to tackle is $f^{(2)}(\gamma)$. But noting that

$$2(e^\gamma + e^{-\gamma}) = 4\left(1 + \frac{\gamma^2}{2!} + \frac{\gamma^4}{4!} + \dots\right) = 4 + \mathcal{O}(\gamma^2)$$

implies that $f^{(2)}(\gamma)\gamma^2 = \gamma^2 + \mathcal{O}(\gamma^4/d)$, which yields the result. \square